International Baccalaureate

MATHEMATICS

Analysis and Approaches (SL and HL)

Lecture Notes

Christos Nikolaidis

TOPIC 4
STATISTICS AND PROBABILITY

Only for HL

JANUARY 2023

4.1 BASIC CONCEPTS OF STATISTICS

In Statistics we deal with data collection, presentation, analysis and interpretation of results. Data can be from

Population	(the entire list of a specified group)
Sample	(a subset of the Population)

We usually investigate a small sample of the population to draw conclusions for the whole population itself.

Numerical data can be

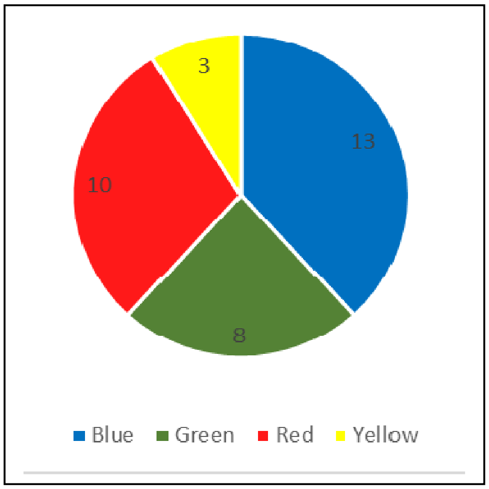
Discrete	OR	**Continuous**
{10,20,30}		[40,100]
{0,1,2,3,…}		R
(finite or numerable set)		(interval)

Data can be organized in several ways. We present some examples below

Frequency table

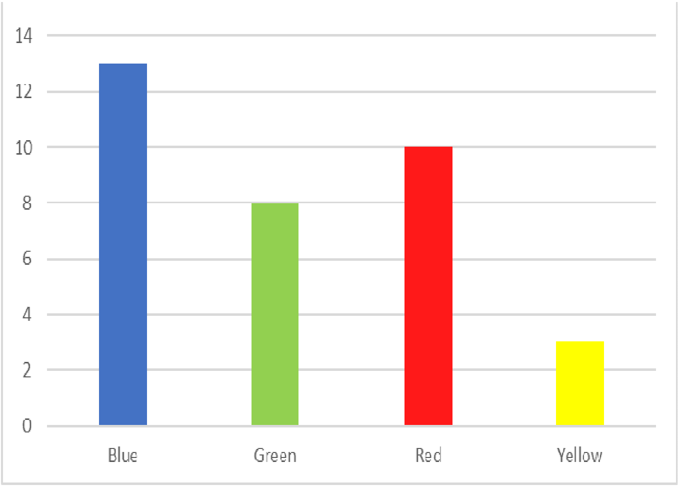
Colored Balls	Frequency
Blue	13
Green	8
Red	10
Yellow	3

Pie chart



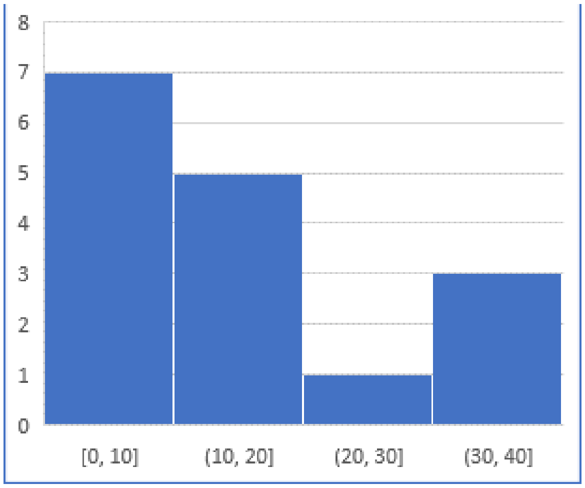
Bar graph (for discrete data)

Colored Balls	Freq
Blue	13
Green	8
Red	10
Yellow	3



Histogram (for continuous data)

Age	Frequency
[0,10)	7
[10,20)	5
[20,30)	1
[30,40)	3



Stem and leaf Diagram

Data
12, 14, 16, 16, 20, 21
21, 21, 25, 32, 39, 40
43, 44, 47, 48, 49, 53

Key: 1|3 represents 13

Stem	Leaf
1	2, 4, 6, 6
2	0, 1, 1, 1, 5
3	2, 9
4	0, 3, 4, 7, 8, 9
5	3

As far as sampling is concerned, it is very crucial to select a sample which is not biased. There are several sampling techniques which face this bias.

Suppose that we have a population of 100,000 people and wish to select a sample of 1000 people. If we select the first 1000 in a list, or the youngest 1000 there is certainly a bias in our selection.

Simple random sampling: We select 1000 people out of a hat
Each member has an equal probability

Systematic sampling: Since 100000/1000=100 (=period) we pick a random starting point (e.g. the 20th person) and pick every 100th person (i.e. 20th, 120th, 220th, …)

Stratified sampling: We divide the population in subgroups (say men and women, or under and over 40 years old). We pick a sample from each group

Quota sampling: As in stratified but we pick proportional samples according to the proportion of the subgroups in the population.

There are advantages and disadvantages in each method. Simple random sampling is fair but it may be very time consuming compared to the systematic sampling. In systematic sample though, if there is a periodic pattern in the population there may be a bias. Suppose that the 100000 are in groups of 100 people. If the first person of the group is the leader, then the sampling method of selecting every 100th person may provide a sample of only leaders or no leaders at all.

4.2 MEASURES OF CENTRAL TENDENCY AND SPREAD

Consider the following numerical data[1]:

10, 20, 20, 20, 30, 30, 40, 50, 70, 70, 80

The total number of entries is **n=11**.

In order to describe these data we use

- 3 measures of central tendency
- 3 measures of spread

The first three measures indicate a representative central value which best describes the data, while the second three measures indicate if our data are very close or dispersed to each other.

♦ MEASURES OF CENTRAL TENDENCY (The 3 M's)

A) MEAN = The sum of all values divided by n.

Here

$$\textbf{mean} = \frac{10+20+20+20+30+30+40+50+70+70+80}{11} = \textbf{40}$$

B) MODE = the most frequent value

Here

mode = 20

C) MEDIAN = The value in the middle

(provided they have been placed in ascending order).

Here, it is the sixth number in the list

median = 30

[1] This set of values is either a **population** or a **sample**.

NOTICE

- For the data 10, 20, 30

  Median = 20

  For the data 10, 20, 30, 40

  Median = 25

  That is, for an **even** number of data,

  median = the mean of the two middle values

- The median is not the $\frac{n}{2}$-th entry as one would possibly expect.

  the median is the $\frac{n+1}{2}$-th entry.

  For example,

  if $n=11$, $\frac{n+1}{2}=6$, thus the median is the 6th entry. See the example above;

  if $n=10$, $\frac{n+1}{2}=5.5$, thus the median is the mean of the 5th and 6th entries; for the 10 entries

  10, 20, 30, 40, 50, 60, 70, 80, 90, 100

  the median is the mean of 50 and 60. Hence **median = 55**

  The median is also denoted by $\mathbf{Q_2}$ (the index 2 will be clarified soon)

- The **mean** is denoted by $\boldsymbol{\mu}$ (or by $\bar{x}$). In fact, we use

  the Greek letter $\boldsymbol{\mu}$ for the whole population.

  the Latin letter $\bar{x}$ for a sample of the population.

  If our data are denoted by $x_1, x_2, \ldots, x_n$, the mean is given by

$$\mu = \frac{x_1 + x_2 + x_3 + \cdots}{n}$$

  or otherwise

$$\mu = \frac{\sum x_i}{n}$$

EXAMPLE 1

Find

a) the integers $a \le b \le c$, given that mean=4, mode=5, median=5.

The median implies that b=5. The mode implies that also c=5.

Then $\frac{a+5+5}{3}=4 \Leftrightarrow a+10=12 \Leftrightarrow a=2$

Therefore, the numbers are 2,5,5.

b) the integers $a \le b \le c \le d$, given that mean=5, mode=7, median=6.

The median implies that either b=c=6 or (b=5 and c=7)

Since the mode is 7 we obtain b=5 and c=d=7.

Then $\frac{a+5+7+7}{4}=5 \Leftrightarrow a+19=20 \Leftrightarrow a=1$

Therefore, the numbers are 1,5,7,7.

♦ MEASURES OF SPREAD

We use the same set of data

10, 20, 20, 20, 30, 30, 40, 50, 70, 70, 80

A) STANDARD DEVIATION

The **standard deviation** is perhaps the most "reliable" measure for spread, as it takes all data into consideration. It measures how far the entries from the mean are. It can be found by using the GDC (directions will be given later on).

The **standard deviation** is denoted[2] either by **σ** or by s_n.

For our example the GDC gives **σ = 22.96.**

[2] In fact,

the Greek letter **σ** is used for the whole population;

the Latin letter s_n is used for a sample of the population

B) RANGE = (maximum value) – (minimum value)

Here

$$\textbf{range} = 80\text{-}10 = \textbf{70}$$

C) INTERQUARTILE RANGE = IQR = $Q_3 - Q_1$

where

Q_1 = LOWER QUARTILE = the median of the values before Q_2

Q_3 = UPPER QUARTILE = the median of the values after Q_2

Here, before the median Q_2=**30**, we have 5 numbers, hence

Q_1=20 (this is the 3rd entry)

Also,

Q_3=70 (it is the 3rd entry from the end)

Therefore,

$$\textbf{IQR} = 70\text{-}20 = \textbf{50}$$

As the estimation of the values Q_1, Q_2, Q_3 is quite tricky, let us see some extra cases in the following example.

EXAMPLE 2 Remember that

- for the value of the median Q_2 we consider the $\frac{n+1}{2}$th entry.
- for the values of Q1 and Q3 we consider only the entries **before** and the entries **after** the median respectively.

a) For n=7 entries: **10, 20, 30, 40, 50, 60, 70**
The median is Q_2=40 (the 4th entry). Hence **Q_1=20, Q_3=60**.

b) For n=8 entries: **10, 20, 30, 40, 50, 60, 70, 80**
The median is Q_2=45 (the 4.5th entry). Hence **Q_1=25, Q_3=65**.

c) For n=9 entries: **10, 20, 30, 40, 50, 60, 70, 80, 90**
The median is Q_2=50 (the 5th entry). Hence **Q_1=25, Q_3=75**.

d) For n=10 entries: **10, 20, 30, 40, 50, 60, 70, 80, 90, 100**
Then Q_2=55 (the 5.5th entry). Hence **Q_1=30, Q_3=80**.

NOTICE

The square of the standard deviation is called **variance**. That is

$$\textbf{variance} = \sigma^2 \quad \textbf{or} \quad s_n^2$$

For our example, $\sigma^2 = 22.96^2 = 527.27$

♦ USE OF GDC

We can use the GDC to easily obtain all these measures.

For Casio CFX we select

- MENU
- STAT
- Complete List 1 with values of x (our data)
- CALC
- (1VAR): We obtain all the statistics.

Notice that

The standard deviation in the GDC is denoted by σ_χ

The variance is not given; it is simply the square of σ_χ

♦ BOX AND WHISKER PLOT

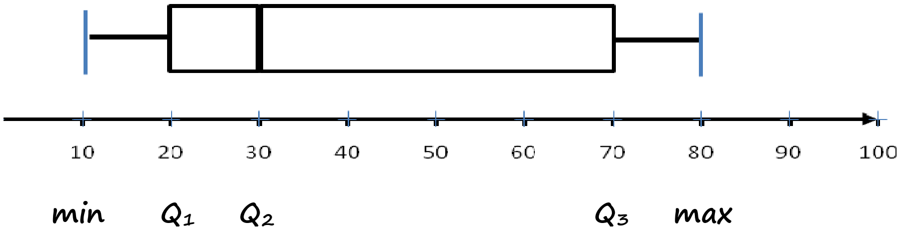
Consider again the initial example

10, 20, 20, 20, 30, 30, 40, 50, 70, 70, 80

In an appropriate horizontal scale we mark 5 figures:

min, Q_1, Q_2, Q_3, max

in the following way:



This diagram is helpful, particularly when we have a large number of entries. It shows the "density" of data within the whole range. In fact, the box plot splits the whole range of data in 4 intervals. Generally speaking, each interval contains 25% of the entries. Thus the following conclusions can be drawn:

The lowest 25% is below Q_1 The upper 25% is above Q_3
The lowest 50% is below Q_2 The upper 50% is above Q_2
The middle 50% is between Q_1 and Q_3

♦ MORE DETAILS

1) Percentiles

The values Q_1, Q_2, Q_3 are also called

Q_1 : 25th-percentile
Q_2 : 50th-percentile
Q_3 : 75th-percentile

Other percentiles may also be defined in a similar way; we will give further examples in the next paragraph.

2) **Outliers**

Very extreme values in a set of data (that is very small or very large) may give a false impression for out data. They are known as outliers. We agree that

an **outlier** is any value

$$\text{below } Q_1 - 1.5\times IQR$$
$$\text{or above } Q_3 + 1.5\times IQR,$$

Such a value is viewed as being too far from the central values to be reasonable. In our example,

$$Q_1 - 1.5\times IQR = 20 - 1.5\times 50 = -55$$

$$Q_3 + 1.5\times IQR = 70 + 1.5\times 50 = 145$$

i.e. there are no outliers.

♦ MORE ON VARIANCE – STANDARD DEVIATION **(only for HL)**

If our data are $x_1, x_2, \ldots, x_n$

the **variance** is given by $\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$

the **standard deviation** is given by $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$

For our example,

$$\textbf{variance} = \frac{(10-40)^2 + (20-40)^2 + (20-40)^2 + \cdots + (80-40)^2}{11} = \mathbf{527.27}$$

$$\textbf{standard deviation} = \sqrt{527.27} = \mathbf{22.96}$$

The variance measures the spread of the data as in fact we find

- the distance of each entry from the mean
- the squares of these distances
- the average of all these square distances

An alternative and more practical formula for the **variance** is given by

$$\sigma^2 = \frac{\sum x_i^2}{n} - \mu^2$$

For our example, we have $\vec{x}=40$ and

$$\frac{\sum x_i^2}{n} = \frac{10^2 + 20^2 + 20^2 + \cdots + 80^2}{11} = \frac{23400}{11} = 2127.27$$

Hence

$$\sigma^2 = 2127.27 - 40^2 = \mathbf{527.27}$$

<u>Proof of the alternative formula</u>

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n} = \frac{\sum (x_i^2 - 2\mu x_i + \mu^2)}{n}$$

$$= \frac{\sum x_i^2}{n} - 2\mu \frac{\sum x_i}{n} + \frac{\sum \mu^2}{n} = \frac{\sum x_i^2}{n} - 2\mu\mu + \frac{n\mu^2}{n}$$

$$= \frac{\sum x_i^2}{n} - 2\mu^2 + \mu^2 = \frac{\sum x_i^2}{n} - \mu^2$$

4.3 FREQUENCY TABLES – GROUPED DATA

Consider again the numerical data:

10, 20, 20, 20, 30, 30, 40, 50, 70, 70, 80

The total number of entries is **n=11**.

An alternative way of presentation is the **frequency table**:

Data **x**	**Frequency** **f**
10	1
20	3
30	2
40	1
50	1
70	2
80	1
	n=11

Let us study again the basic measures for these data.

- MEASURES OF CENTRAL TENDENCY (The 3 M's)

A) MEAN = The sum of all values divided by n.

The MEAN is given by

$$\textbf{mean} = \frac{1\times10+3\times20+2\times30+1\times40+1\times50+2\times70+1\times80}{11} = \mathbf{40}$$

In general, given that f_i is the frequency of the entry x_i, the formula is

$$\mu=\frac{f_1x_1+f_2x_2+f_3x_3+\cdots}{n} \qquad \text{or otherwise} \qquad \mu=\frac{\sum f_ix_i}{n}$$

B) MODE = the most frequent value

It is very obvious now. The entry x of the highest frequency is

mode = 20

C) MEDIAN = The value in the middle

It is still the entry in position $\frac{n+1}{2}$, that is the 6th entry.

We can easily see that this is 30.

It helps here to add an extra column in the table above with the so-called **cumulative frequencies:**

Data x	Frequency f	Cumulative frequency (c.f.)
10	1	1
20	3	4
30	2	6
40	1	7
50	1	8
70	2	10
80	1	11
	n=11	

It simply gives the total number of entries up to each row. For example, the total number of entries up to 20 is 1+3=4.
The MEDIAN, i.e. the 6th entry, is 30.

- MEASURES OF SPREAD

A) STANDARD DEVIATION

Again, it can be directly obtained by the GDC.

For our example the GDC gives **σ = 22.96.**

Thus the **variance** is **σ^2 = 527.27**

B) RANGE = (maximum value of x) – (minimum value of x)

It is very obvious here

range = 80-10 = 70

C) INTERQUARTILE RANGE = IQR = $Q_3 - Q_1$

The cumulative frequency table helps here as well.

The median Q_2=**30** is in the 6th position.

Thus, before the median we have 5 entries. Since $\frac{n+1}{2}=3$,

Q_1=20 (this is the 3rd entry)

and

Q_3=70 (this is the 3rd entry from the end)

Therefore,

IQR = 70-20 = 50

♦ USE OF GDC

We can use the GDC to easily obtain all these measures.
For Casio CFX we select

- MENU
- STAT
- Complete List 1 with values of x (our data)
  List 2 with frequencies
- CALC
- SET: we check the first two lines
  The first line is OK. (1Var XList :List1)
  For the second line (1Var Freq :----), select between
  F1: enter 1, if there are no frequencies
  F2: enter List 2 to consider frequencies
- Go back (EXIT)
- 1VAR: We obtain all the statistics.

  Check the value of **n** first (number of entries), to ensure that all data have been considered.

NOTICE (for the GDC)

- The variance is not given; it is simply the square of σ_χ
- Since the GDC gives **minX,Q1,Med,Q3,maxX** remember that

  Range = **maxX − minx** Interquartile Range = **Q3 − Q1**

  The box and whisker plot uses exactly those 5 measures
- Extra information given:

  Σx : the sum of all entries, i.e. $x_1+x_2+x_3+\ldots$

  Σx²: the sum of the squares, i.e. $x_1^2+x_2^2+x_3^2+\ldots$

  s_χ : it is known as unbiased st. deviation (not in the syllabus!)

♦ GROUPED DATA

Suppose that 100 students took an exam and obtained scores from 1 to 60 (full marks), according to the following table:

Score (x)	Midpoint (for x)	No of students (frequency f)	Cumulative frequency (cf)
$0<x\leq10$	5	8	8
$10<x\leq20$	15	12	20
$20<x\leq30$	25	10	30
$30<x\leq40$	35	25	55
$40<x\leq50$	45	35	90
$50<x\leq60$	55	10	100
		n=100	

i.e. 8 students obtained scores from 1 up to 10, and so on.

- The mean and the standard deviation are still calculated as in a usual frequency table, but now $x_1,x_2,x_3,\ldots$ are the midpoints of the intervals.

  For example,

$$\mu=\frac{8\times5+12\times15+10\times25+25\times35+35\times45+10\times55}{100}=34.7$$

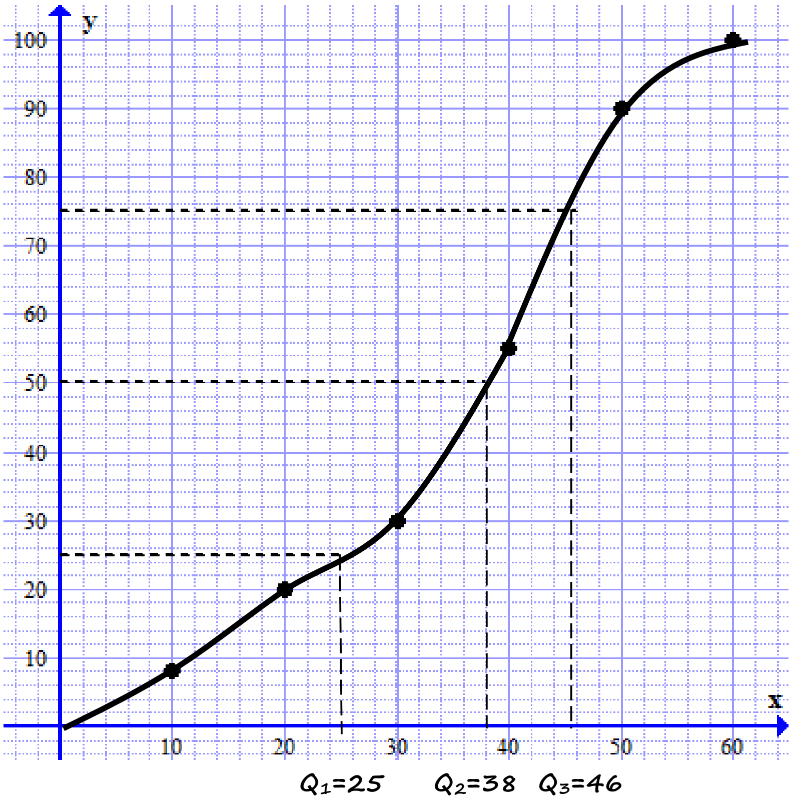
These measures may also be obtained by the GDC, where the LIST1 contains the midpoints of x. Here,

$$\mu=34.7 \qquad \sigma=14.31$$

- Moreover, instead of the mode we have the **modal group** here. That is the interval of the highest frequency. In our example, the modal group is $40<x\leq 50$.

- For the median Q2 and the quartiles Q1 and Q3:
  we need to draw the so-called **cumulative frequency diagram**
  x-axis: values of x (we consider upper bounds of intervals)
  y-axis: cumulative frequencies

x: up to	≤10	≤20	≤30	≤40	≤50	≤60
y: c.f	8	20	30	55	90	100



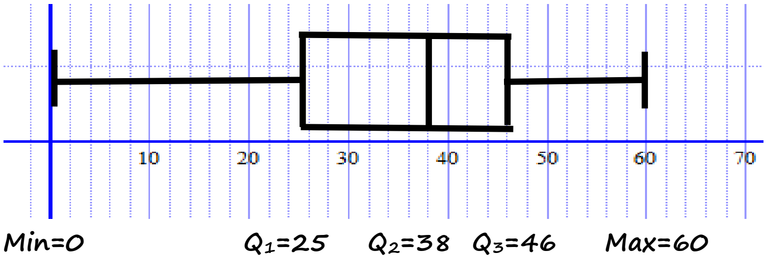
For the estimation of Q_1, Q_2, Q_3 follow

Step 1: Divide y-axis into four equal parts
(Here we divide at y=25, y=50, y=75)

Step 2: Draw three horizontal lines until you meet the curve

Step 3: Draw three vertical lines from the intersection points
Obtain Q_1, Q_2, Q_3 on x-axis (look at above)

Below that graph we can easily draw box and whisker plot:



- Remember that the values Q_1, Q_2, Q_3 are also called

  Q_1 : 25th-percentile

  Q_2 : 50th-percentile

  Q_3 : 75th-percentile

In the same way we can find any **percentile**. For example, for the 40th-percentile

Estimate 40% of n: here 40% of 100 students is 40;

Draw a horizontal line at y=40 until you meet the curve;

Then draw a vertical line;

Hence

40th-percentile = 35.

In other words, 40% of the students have scores below 35.

- Let us check if there are **outliers**:

  IQR = 46-25=21

$$Q_1 - 1.5\times IQR = 25 - 1.5\times 21 = - 6.5$$

$$Q_3 + 1.5\times IQR = 46 + 1.5\times 21 = 77.5$$

There are no scores lower than -6.5 or greater than 77.5, that is there are no outliers.

- MORE ON VARIANCE – STANDARD DEVIATION (only for HL)

  Given that f_i is the frequency of the entry x_i, the formulas now become:

$$\sigma^2 = \frac{\sum f_i (x_i - \mu)^2}{n}$$

  thus

$$\sigma = \sqrt{\frac{\sum f_i (x_i - \mu)^2}{n}}$$

  In our example,

$$\textbf{variance} = \frac{1\times(10-40)^2 + 3\times(20-40)^2 + 2\times(30-40)^2 + \cdots}{11} = \mathbf{527.27}$$

  and then

$$\textbf{standard deviation} = \sqrt{527.27} = \mathbf{22.96}$$

The alternative formula for the **variance** takes the form

$$s_n{}^2 = \frac{\sum f_i x_i{}^2}{n} - \vec{x}^2$$

For our example, we have $\vec{x}=40$ and

$$\frac{\sum f_i x_i{}^2}{n} = \frac{1\times 10^2 + 3\times 20^2 + 2\times 20^2 + \cdots}{11} = \frac{23400}{11} = 2127.27$$

Hence

$$\sigma^2 = 2127.27 - 40^2 = \mathbf{527.27}$$

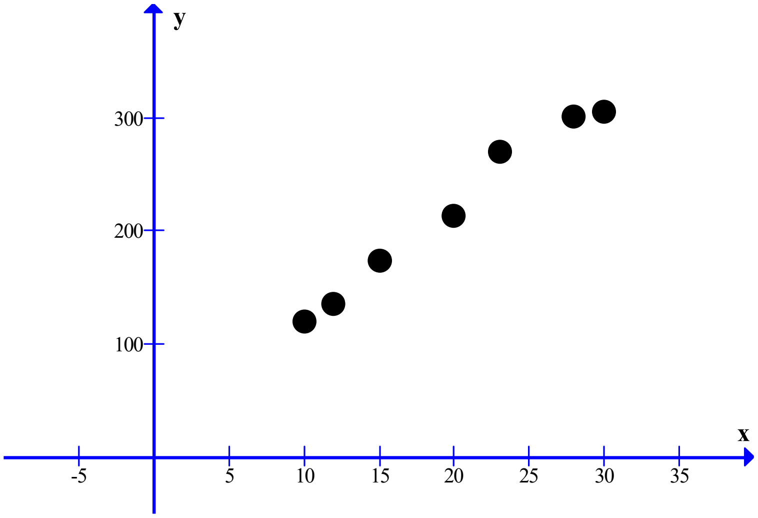
(The proof of the alternative formula is very similar to what we have seen in 5.1)

4.4 REGRESSION

We have a list of paired data. For example

x	10	12	15	20	23	28	30
y	120	135	174	213	270	301	305

We assume that x is the *independent variable*, y is the *dependent variable*. Let us also see these points (x,y) on a **scatter diagram**.



The main question here is whether there is a linear relationship between the values of x and the corresponding values of y.

There is a parameter **r**, called **correlation coefficient**[3] that gives the extent of this relationship. It takes values

$$-1 \le r \le 1$$

The closest to the ends ±1, the more our data are linearly related.
(–1 implies a negative slope while +1 implies a positive slope)
The closest to 0, the less our data are linearly related.

There is also a line **y=ax+b** that best fits our data; it is known as ***regression line***. We can easily obtain these details by using a GDC.

[3] It is known as Pearson's product-moment correlation coefficient

♦ USE OF GDC

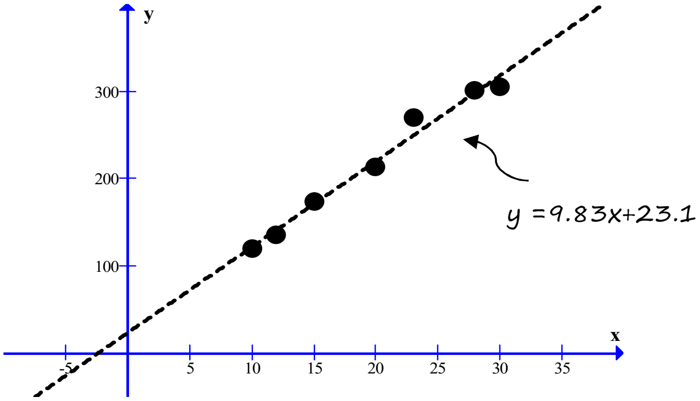
For Casio CFX we select

- MENU
- STAT
- Complete List 1 with values of x; List 2 with values of y
- CALC
- REG
- X
- aX+b : look at the values of a,b,r.

For our example,

r =0.99	there is a very strong correlation between x and y

a =9.83
b =23.1
	The regression line is **y =9.83x+23.1**



By using the regression line y=f(x) we may predict values of y corresponding to values of x that are not in the list. For example

for x=18, we estimate $y = 9.83\times18+23.1 \cong 200$

for x=40, we estimate $y = 9.83\times40+23.1 \cong 416$

Notice that x=18 is within the range of our list while x=40 is not. f(18)=200 is known as **interpolation,** f(40)=416 as **extrapolation.** In general, interpolations are more reliable than extrapolations.

Notice. In order to predict a value of x corresponding to a given y we do not use the same regression line. We find a new regression line for x on y. In our example, the GDC gives **x =0.0997y–1.92**

♦ CHARACTERISTICS OF THE REGRESSION LINE $y=ax+b$

The regression line

- passes through the point M($\bar{x}$,$\bar{y}$), where

  $\bar{x}$ = the mean of the values of x

  $\bar{y}$ = the mean of the values of y

- separates the points in (almost) two halves: half of the points are above and half below the line.

The values of $\bar{x}$,$\bar{y}$ can also be obtained by the GDC (together with other statistics). In the STAT mode, after inserting the values of x and y, select

- CALC
- 2VAR: We obtain all the statistics, separately for x's and y's

In our example

$$\bar{x}= 19.7 \qquad \bar{y}=216.9$$

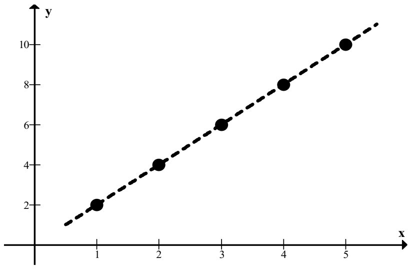
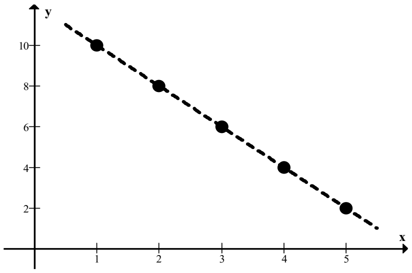
Thus the line passes through the point M(19.7, 216.9).

♦ CHARACTERISTICS OF THE CORRELATION COEFFICIENT r

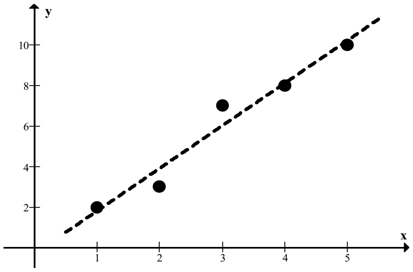
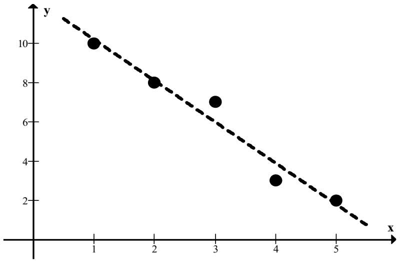
The correlation between x and y is characterised according to the value of r as follows:

-1 -0.75	-0.5	-0.25	0 0.25	0.5	0.75	1
strong negative correlation	**moderate negative correlation**	**weak negative correlation**	**very weak or no correlation**	**weak positive correlation**	**moderate positive correlation**	**strong positive correlation**

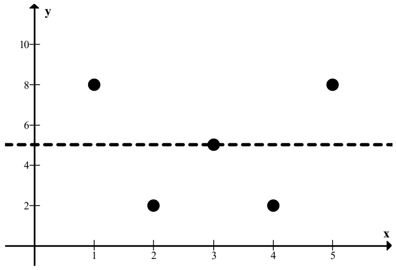
To better understand the correlation coefficient r, let us see some characteristic cases (find the results below in your GDC for practice).

<table>
<tr><th>Data</th><th>Scatter diagram</th><th>Results</th></tr>
<tr><td><table><tr><th>x</th><th>y</th></tr><tr><td>1</td><td>2</td></tr><tr><td>2</td><td>4</td></tr><tr><td>3</td><td>6</td></tr><tr><td>4</td><td>8</td></tr><tr><td>5</td><td>10</td></tr></table></td><td>
</td><td><b>r =1</b>
perfect positive
correlation
<b>Regression line:</b>
<b>y=2x</b></td></tr>
<tr><td><table><tr><th>x</th><th>y</th></tr><tr><td>1</td><td>10</td></tr><tr><td>2</td><td>8</td></tr><tr><td>3</td><td>6</td></tr><tr><td>4</td><td>4</td></tr><tr><td>5</td><td>2</td></tr></table></td><td>
</td><td><b>r =-1</b>
perfect negative
correlation
<b>Regression line:</b>
<b>y=-2x+12</b></td></tr>
</table>

Let us slightly modify our data

<table>
<tr><td><table><tr><th>x</th><th>y</th></tr><tr><td>1</td><td>2</td></tr><tr><td>2</td><td>3</td></tr><tr><td>3</td><td>7</td></tr><tr><td>4</td><td>8</td></tr><tr><td>5</td><td>10</td></tr></table></td><td>
</td><td><b>r =0.98</b>
strong positive
correlation
<b>Regression line:</b>
<b>y=2.1x-0.3</b></td></tr>
<tr><td><table><tr><th>x</th><th>y</th></tr><tr><td>1</td><td>10</td></tr><tr><td>2</td><td>8</td></tr><tr><td>3</td><td>7</td></tr><tr><td>4</td><td>3</td></tr><tr><td>5</td><td>2</td></tr></table></td><td>
</td><td><b>r =-0.98</b>
strong negative
correlation
<b>Regression line:</b>
<b>y=-2.1x+12.3</b></td></tr>
</table>

and a final extreme case

<table>
<tr><td><table><tr><th>x</th><th>y</th></tr><tr><td>1</td><td>8</td></tr><tr><td>2</td><td>2</td></tr><tr><td>3</td><td>5</td></tr><tr><td>4</td><td>2</td></tr><tr><td>5</td><td>8</td></tr></table></td><td>
</td><td><b>r = 0</b>
no correlation
at all
<b>Regression line:</b>
<b>y=5</b></td></tr>
</table>

4.5 ELEMENTARY SET THEORY

♦ BASIC NOTIONS

In elementary set theory, a **set** is just a collection of objects (or **elements**). It is usually denoted by a capital letter. For example,

R = the set of real numbers
Q = the set of rational numbers

When listed, the **elements** of a set are separated by commas "," and included between the symbols { and }. For example,

$N = \{0,1,2,3,4,\ldots\}$ (i.e. the set of natural numbers)
$Z = \{\ldots,-3,-2,-1,0,1,2,3,\ldots\}$ (i.e. the set of all integers)

or less popular sets, such as

$A = \{1,2,3\}$ (it contains only 3 elements)
$B = \{a,b,c,d\}$ (it contains 4 letters)
C = {Chris, Mary, Tom} (it contains 3 names)
etc

To declare that <u>the element a is contained in set B</u> we write

$$a \in B$$

To declare that <u>the element f is not contained in set B</u> we write

$$f \notin B$$

The most trivial set is the **empty set**. It contains no elements, it is denoted by { } or by the symbol $\varnothing$.

Let us consider the set $A = \{1,2,3\}$. The **subsets** of A are sets that contain some (or none or all) elements of A. There are 8 subsets:

$\varnothing$
$\{1\}$, $\{2\}$, $\{3\}$
$\{1,2\}$, $\{1,3\}$, $\{2,3\}$
$\{1,2,3\}$

In general,

if A contains **n** elements, there are $\mathbf{2^n}$ subsets.

Indeed, here, A contains 3 elements and possesses $2^3=8$ subsets.

If A = {1,2,3} and B = {1,2}, to declare that B is a subset of A, we write

$$B \subseteq A$$

Do not forget that always

$\varnothing \subseteq A$ (The empty set is a subset of any set)

$A \subseteq A$ (Any set is a subset of itself)

All subsets of A except itself are also called **proper subsets**. To emphasize that B is a proper subset of A we write

$$B \subset A$$

♦ VENN DIAGRAMS

We usually refer to a large set S, called **universal set**, and consider several subsets of S.

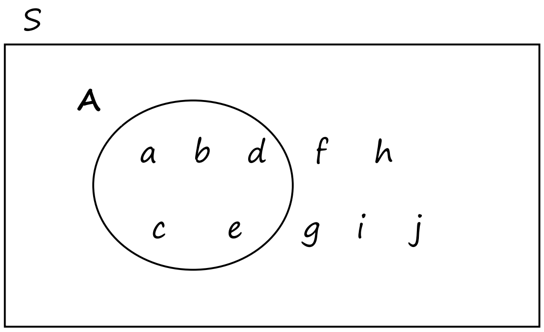
Let

$$S = \{ a,b,c,d,e,f,g,h,i,j \}$$

be our universal set. We consider the subset

$$A = \{a,b,c,d,e\}$$

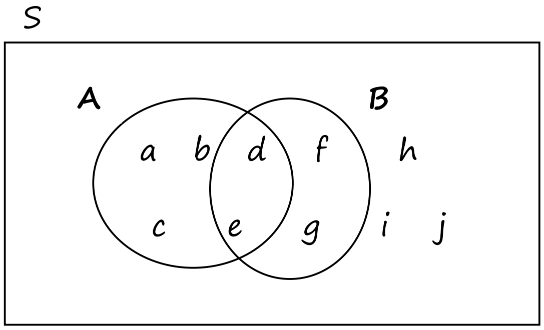
A helpful way to present this information is by using a **Venn diagram**:



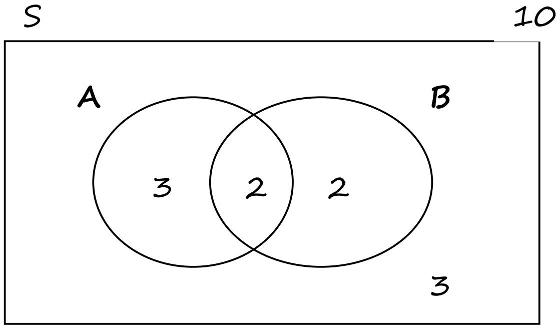
If we also consider the subset

$$B = \{d,e,f,g\}$$

the Venn diagram becomes



As we usually deal with large universal sets, in a Venn diagram we are not interested so much for the elements themselves but only for the number of elements in each region. In this case the Venn diagram above takes the form



We denote by

n(A) = the number of elements of set A

In our example

$$n(S) = 10$$

$$n(A)=5 \qquad n(B)=4$$

Notice that the number n(A)=5 does not appear on the Venn diagram. The subset A consists of two regions of size 3 and 2, thus

$$n(A)=3+2=5$$

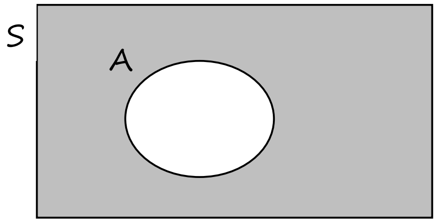
Now we can study some basic operations between sets. Let us refer again to our example where S = { a,b,c,d,e,f,g,h,i,j } and

A = {a,b,c,d,e}

B = {d,e,f,g}

- THE COMPLEMENT OF A: A′ **(not A)**

It contains the elements that are not in A.

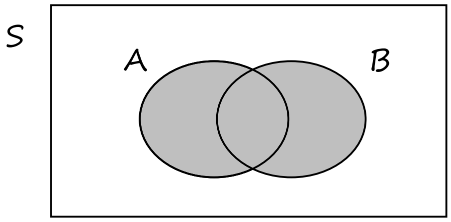


In our example A′ = {f,g,h,i,j}

Sometimes the complement of A is also denoted by $\overline{A}$.

- THE UNION OF A AND B: A∪B **(A or B)**

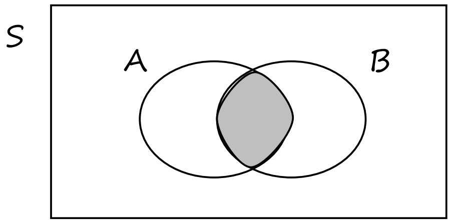
It contains all the elements that are either in A or in B.



In our example A∪B = {a,b,c,d,e,f,g}

- THE INTERSECTION OF A AND B: A∩B **(A and B)**

It contains the common elements of A and B.



In our example A∩B = {d,e}

- A BASIC PREPERTY

$$n(A\cup B) = n(A) + n(B) - n(A\cap B)$$

Indeed, in our example

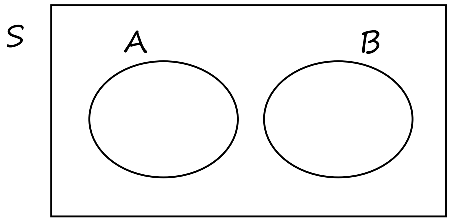
$$n(A\cup B)=7,\ n(A)=5,\ n(B)=4,\ n(A\cap B)=2$$

Notice that $A\cup B$ contains 7 elements, not 5+4=9, as in n(A)+n(B) we count the common elements twice. Thus,

$$7 = 5 + 4 - 2$$

- MUTUALLY EXCLUSIVE SETS

If $A\cap B=\varnothing$, then $n(A\cap B)=0$



In this case only

$$n(A\cup B) = n(A) + n(B)$$

and the two sets A and B are said to be **mutually exclusive**.

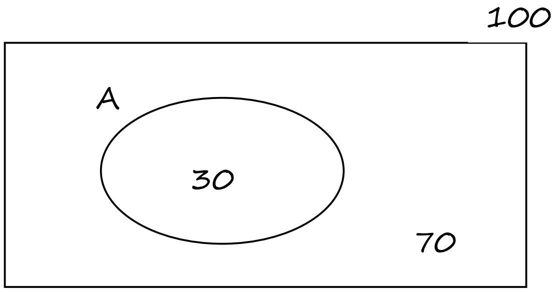
4.6 PROBABILITY

We start again with a universal set S. In probability theory this set is known as the **sample space**; it contains all possible outcomes of a game, or experiment, etc. The subsets A,B, … of the sample space S are called **events**.

Consider the sample space S. The number of elements in S, that is n(S), is denoted by TOTAL. The probability of some event A is simply defined by

$$P(A)=\frac{n(A)}{TOTAL}$$

For example, in the following Venn diagram, the sample space S contains 100 elements, while the event A contains 30 elements



$$P(A)=\frac{n(A)}{TOTAL}=\frac{30}{100}=0.3$$

In simple words, if we choose an element from S at random, (provided that every element is equally likely to be selected), the probability that this element belongs to A is 30 out of 100, otherwise 30% (that is 0.3).

We understand that

$$0\le P(A)\le 1$$

Clearly

$$P(\varnothing)=0 \text{ and } P(S) = 1$$

- COMPLEMENTARY EVENTS

In our example above P(A′) = 0.7

In general

$$P(A') = 1 - P(A)$$

- COMBINED EVENTS

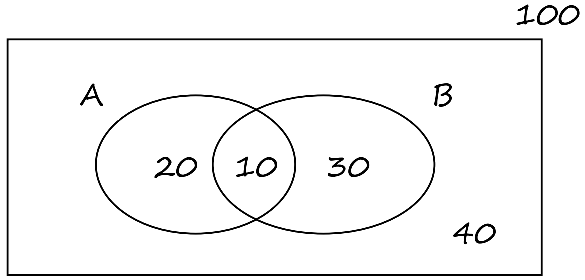
Remember the basic property for combined events

$$n(A \cup B) = n(A) + n(B) - n(A \cap B)$$

If we divide all terms by the TOTAL we obtain

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

For example, consider



Then

$$P(A) = 0.3,\ P(B) = 0.4,\ P(A')=0.7,\ P(B')=0.6$$

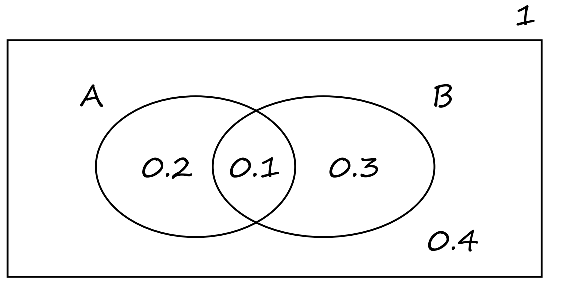
Also

$$P(A \cap B)=0.1,\ P(A \cup B)=0.6$$

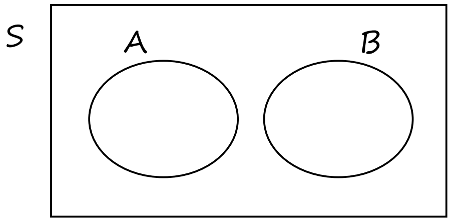
Clearly

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$0.6 = 0.3 + 0.4 - 0.1$$

A Venn diagram may also contain probabilities instead of numbers of elements. The Venn diagram above takes the form



- MUTUALLY EXCLUSIVE SETS



We have seen that two events are **mutually exclusive** if

$A\cap B=\varnothing$ or equivalently $n(A\cap B)=0$;

Equivalently if

$$P(A\cap B)=0$$

In this case only

$$P(A\cup B) = P(A) + P(B)$$

EXAMPLE 1

Given that P(A) = 0.5, P(B) = 0.3, P(A∪B)=0.6, let us construct a Venn diagram representing the combined events A and B

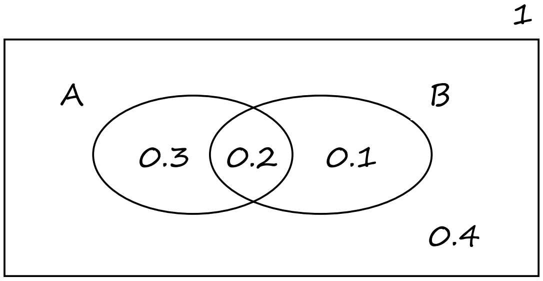
Notice that

$$P(A\cup B)\neq P(A)+P(B)$$

$$0.6 \neq 0.8$$

The difference implies the existence of an intersection; P(A∩B)=0.2

Starting from the intersection 0.2 we may easily complete the following Venn diagram



After completing the Venn diagram, we are in a position to answer any probability question. For example

$P(A\cap B') = 0.3$	$P(A'\cap B) = 0.1$	$P(A'\cap B') = 0.4$
$P(A\cup B') = 0.9$	$P(A'\cup B) = 0.7$	$P(A'\cup B') = 0.8$

♦ TABLES

Another way to represent sets in order to find probabilities is the tabular form below. It is appropriate when the sample space is partitioned in disjoint subsets according two different criteria; for example MALE-FEMALE and SMOKERS-NON SMOKERS.

Let us consider the following group of 200 people

	male	**female**	**Total**
smoker	40	20	60
non-smoker	80	60	140
Total	120	80	200

In order to find the probability of a group (or combination of groups) we simply divide its size by 200, the total number of people. Thus

If we select a person at random the probability that this person is

- **male** is $P(\text{male})=\frac{120}{200}=0.6$
- **female** is $P(\text{female})=\frac{80}{200}=0.4$
- **smoker** is $P(\text{smoker})=\frac{60}{200}=0.3$
- **non-smoker** is $P(\text{non}-\text{smoker})=\frac{140}{200}=0.7$
- **male** AND **smoker** $P(\text{male}\cap\text{smoker})=\frac{40}{200}=0.2$
- **male** OR **smoker** $P(\text{male}\cup\text{smoker})=\frac{140}{200}=0.7$

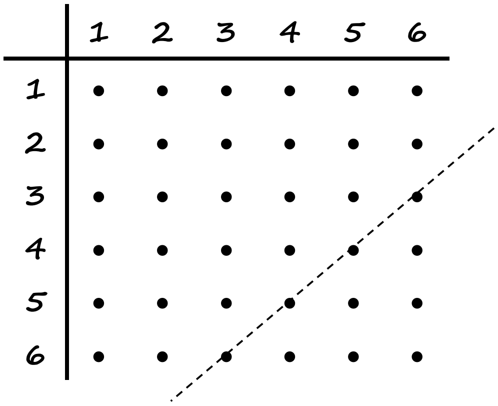
Notice: In the last probability, we consider the column of male and the row of smokers, but the combination male-smoker is counted only once. It holds again

$$P(\text{male}\cup\text{smoker})=P(\text{male})+P(\text{smoker})-P(\text{male}\cap\text{smoker})$$

Some problems require particular techniques for counting the appropriate group size. Tossing two dice is a characteristic example.

♦ TWO DICE

We toss two dice. There are 36 possible outcomes (combinations of scores). The following table helps to visualize the outcomes



	1	2	3	4	5	6
1	●	●	●	●	●	●
2	●	●	●	●	●	●
3	●	●	●	●	●	●
4	●	●	●	●	●	●
5	●	●	●	●	●	●
6	●	●	●	●	●	●

Notice that there is only one combination **ones** (the first dot; 1-1) but two combinations of **one-two** (1-2 and 2-1).

We find the following probabilities:

$$P(\text{two sixes}) = \frac{1}{36}$$ (the very last dot)

$$P(\text{at leat one six}) = \frac{11}{36}$$ (last column and last row)

$$P(\text{exactly one six}) = \frac{10}{36}$$ (why?)

$$P(\text{same score}) = \frac{6}{36}$$ (the main diagonal: 1-1, etc)

$$P(\text{sum of scores} = 9) = \frac{4}{36}$$ (the dotted line)

$$P(\text{sum of scores} > 9) = \frac{6}{36}$$ (below the dotted line)

$$P(\text{sum of scores} < 9) = \frac{26}{36}$$ (above the dotted line)

4.7 CONDITIONAL PROBABILITY – INDEPENDENT EVENTS

Notice the following difference in notation

$P(A)$	means "probability of A"
$P(A\|B)$	means "probability of A, given B"

Intuitively, we expect that

"the probability that it will rain in some day"

is different than

"the probability that it will rain in some day,
given that this is a day of September"

In a more mathematical example, suppose that we pick a whole number in the range 1-100. Let

A = "we pick 17"

Clearly $P(A)=\frac{1}{100}$.

However, if we know the information

B = "the number selected has two digits"

then $P(A|B)=\frac{1}{90}$ (there are 90 two-digit numbers)

♦ FORMAL DEFINITION OF $P(A|B)$

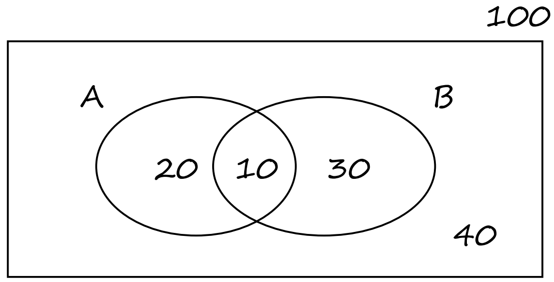
The conditional probability is given by the formula

$$P(A|B)=\frac{n(A\cap B)}{n(B)} \quad \text{or} \quad P(A|B)=\frac{P(A\cap B)}{P(B)}$$

We will clarify the definition by using Venn diagrams and Tables

- $P(A|B)$ IN A VENN DIAGRAM

Let us consider the example



We know that $P(A)=\frac{30}{100}$. What about $P(A|B)$?

We start with the given event B; now the total number is not 100, the size of the whole sample space, but only 40, the size of B:

$$P(A|B)=\frac{?}{40} \quad \leftarrow \text{given B}$$

How many elements of A are inside the given space B? Only 10. Therefore,

$$P(A|B)=\frac{10}{40}$$

NOTICE

In fact, in the last result we apply the formula

$$P(A|B)=\frac{n(A\cap B)}{n(B)}=\frac{10}{40}=0.25$$

If we divide both the numerator and the denominator by the TOTAL number of the sample space we obtain the formal definition

$$P(A|B)=\frac{P(A\cap B)}{P(B)}=\frac{\frac{10}{100}}{\frac{40}{100}}=0.25$$

- Similarly we obtain

$$P(B|A)=\frac{10}{30} \quad \leftarrow \text{given A}$$

- Similarly we obtain

$$P(A'|B)=\frac{30}{40} \qquad P(A|B')=\frac{20}{60} \qquad P(A'|B')=\frac{40}{60}$$

♦ P(A|B) IN A TABLE

Perhaps it is much easier to observe the conditional probability in tables. Consider again the example

	male	**female**	**Total**
smoker	40	20	60
non-smoker	80	60	140
Total	120	80	200

Observe the difference between the probalitites

P(**smoker**) the person is a smoker

P(**smoker**|**male**) the person is a smoker given it is male

Clearly,

$$P(\textbf{smoker}) = \frac{60}{200}$$

$$P(\textbf{smoker}|\textbf{male}) = \frac{40}{120} \quad \leftarrow \text{given male}$$

- Similarly we obtain

$$P(\textbf{male}|\textbf{smoker})=\frac{40}{60} \quad \leftarrow \text{given smoker}$$

- Similarly we obtain

$$P(\textbf{female}|\textbf{smoker})=\frac{20}{60}\approx 0.33 \qquad P(\textbf{non-smoker}|\textbf{female})=\frac{60}{80}=0.75$$

♦ INDEPENDENT EVENTS

The events A and B are said to be ***independent*** if

$$P(A|B) = P(A)$$

In other words, the event B does not affect A;
the probability of A remains the same, either B is given or not!

- Similarly, in this case it holds $P(B|A) = P(B)$

That is, the event A does not affect B.

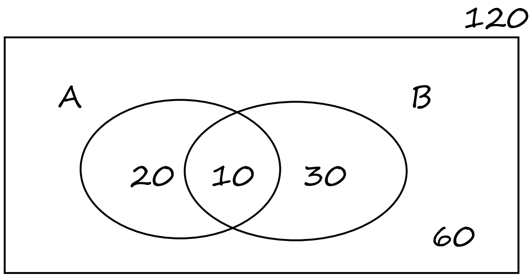
- In this case the definition $P(A|B)=\dfrac{P(A\cap B)}{P(B)}$ gives

$$P(A\cap B)=P(A|B)\cdot P(B) \Rightarrow P(A\cap B)=P(A)\cdot P(B)$$

To summarize

A and B are ***independent***	$P(A\|B) = P(A)$ (1)
	$P(B\|A) = P(B)$ (2)
	$P(A\cap B)=P(A)\cdot P(B)$ (3)

EXAMPLE 1



We can show in three different ways that A and B are independent

- $P(A) = \dfrac{30}{120} = \dfrac{1}{4}$ and $P(A|B) = \dfrac{10}{40} = \dfrac{1}{4}$ thus (1) holds
- $P(B) = \dfrac{40}{120} = \dfrac{1}{3}$ and $P(B|A) = \dfrac{10}{30} = \dfrac{1}{3}$ thus (2) holds
- $P(A\cap B)=\dfrac{10}{120}=\dfrac{1}{12}$ $P(A)\cdot P(B)=\dfrac{1}{4}\cdot\dfrac{1}{3}=\dfrac{1}{12}$ thus (3) holds

NOTICE

- Many students confuse the terms

  Mutually exclusive events and **Independent events**

  Remember

  Mutually exclusive events means $A\cap B=\varnothing$

  Independent events means $P(A\cap B)=P(A)\cdot P(B)$

- Mind that

  $P(A\cup B) = P(A) + P(B) - P(A\cap B)$ holds in general

  $P(A\cap B)=P(A)\cdot P(B)$ holds for independent events

  In particular for independent events, it is sometimes useful to combine these two formulas in the following one

  $$P(A\cup B) = P(A) + P(B) - P(A)\cdot P(B)$$

- Sometimes we know beforehand that two events are independent. Thus, for their combination we can apply the formula $P(A\cap B)=P(A)\cdot P(B)$

  For example,

  we toss a die and a coin; Find the probability that the die shows a SIX and the coin shows a HEAD.

  We call

  A = "the die shows a SIX" $P(A) = \frac{1}{6}$

  B = "the coin shows a HEAD" $P(B) = \frac{1}{2}$

  The events A and B are clearly independent and for their combination it holds

  $$P(A\cap B)=P(A)\cdot P(B)=\frac{1}{6}\frac{1}{2}=\frac{1}{12}$$

EXAMPLE 2

Let P(A)=0.4 and P(B)=0.3. Find $P(A\cup B)$ in the following cases

a) A and B are mutually exclusive

b) A and B are independent

c) $P(A\cap B)=0.2$

d) $P(A|B)=0.2$

Solution

a) $P(A\cup B) = P(A) + P(B) = 0.4 + 0.3 = 0.7$

b) $P(A\cup B) = P(A) + P(B) - P(A)\cdot P(B) = 0.4+0.3-(0.4)(0.3) = 0.58$

c) $P(A\cup B) = P(A) + P(B) - P(A\cap B) = 0.4+0.3-0.2 = 0.5$

d) $P(A\,|\,B)=\dfrac{P(A\cap B)}{P(B)} \Rightarrow P(A\cap B)= P(A|B)P(B)= (0.2)(0.3) = 0.06$

Hence, $P(A\cup B) = P(A)+P(B)-P(A\cap B) = 0.4+0.3-0.06 = 0.64$

EXAMPLE 3

Let A and B be independent events with

$$P(A)=0.4 \quad \text{and} \quad P(A\cup B)=0.7.$$

Find P(B).

Solution

For independent events it holds

$$P(A\cup B) = P(A) + P(B) - P(A)\cdot P(B)$$
$$\Leftrightarrow 0.7 = 0.4 + P(B) - 0.4P(B)$$
$$\Leftrightarrow 0.3 = 0.6P(B)$$
$$\Leftrightarrow P(B)= 0.5$$

4.8 TREE DIAGRAMS

Very often we have to estimate the probability in a sequence of events under different scenarios. The best way to represent such a problem is by a **tree diagram.**

PROBLEM 1. We play a game with two possible results.

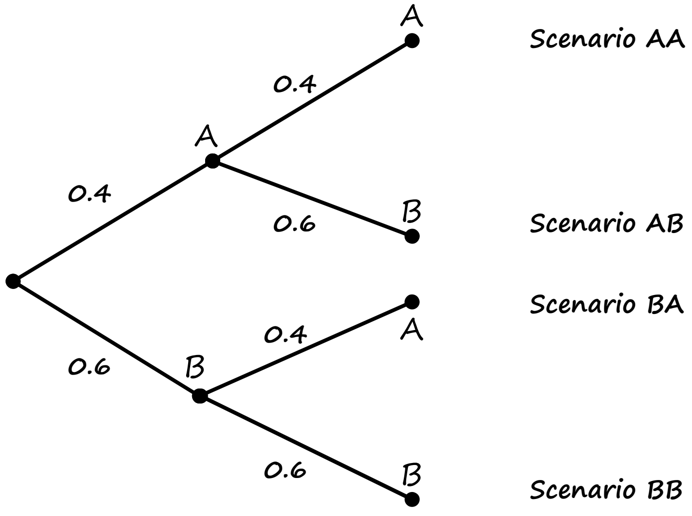
For example we pick one of the following 10 letters

AAAA BBBBBB

The results are

A	with probability 0.4
B	with probability 0.6

We play the game twice. All possible scenarios are shown below; the corresponding probabilities are shown on the branches of the tree:



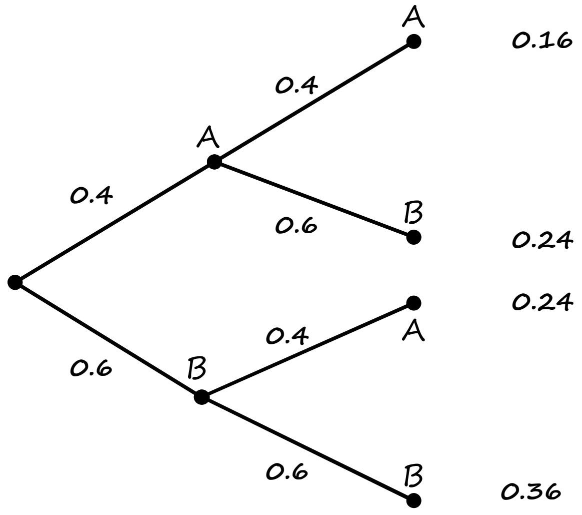
Next, for each scenario we multiply the corresponding probabilities

for **AA**: (0.4)^x(0.4) = 0.16

for **AB**: (0.4)^x(0.6) = 0.24,

etc

Thus, the final "picture" of the tree diagram is as follows



(notice that the sum of the resulting probabilities is 1).

Now any probability may be found by adding the relevant results.

Namely, the probability

- to obtain two A's is **0.16**
- to obtain two B's is **0.36**
- to obtain first A and then B is **0.24**
- to obtain one A, one B is 0.24 + 0.24 = **0.48**
- to obtain the same result is 0.16 + 0.36 = **0.52**

  (thus to obtain different results is 1- 0.52 = **0.48**)

If we refer to the number of A's, the probability

- to obtain <u>no</u> A is **0.36**
- to obtain <u>exactly one</u> A is 0.24 + 0.24 = **0.48**
- to obtain <u>at least one</u> A is 0.24 + 0.24 + 0.16 = **0.64**
- to obtain <u>at most one</u> A is 0.24 + 0.24 + 0.36 = **0.84**

PROBLEM 2. We play again the previous game once. According to the result we play a different second game.

For example we pick one of the following 10 letters

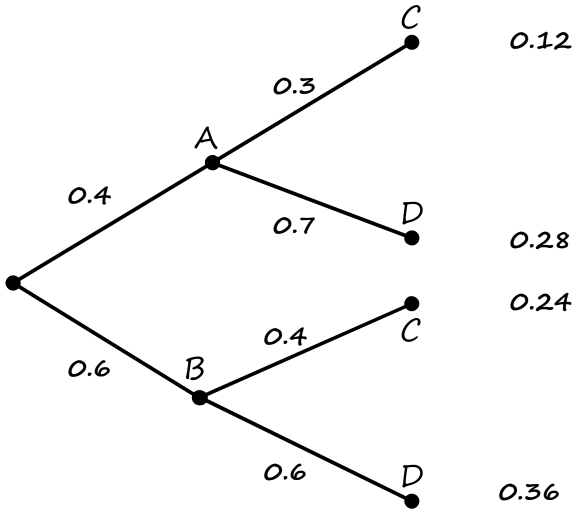
AAAA BBBBBB

If the first result is A we pick one letter among **CCC DDDDDDD**

If the first result is B we pick one letter among **CCCC DDDDDD**

What is the probability to obtain C?

A tree diagram is particularly helpful in such a situation where the second game depends on the first one:



(notice again that the sum of the resulting probabilities is 1).

Thus

the probability to obtain C is 0.12 + 0.24 = **0.36**

the probability to obtain D is 0.28 + 0.36 = **0.64**

It is worthwhile to mention the following probabilities:

- to obtain A **and** C. It is **0.12**
  It is in fact $P(A\cap C)$ and refers to the first scenario
- to obtain A **or** C. It is 0.12 + 0.28 + 0.24 = **0.64**
  It is in fact $P(A\cup C)$ and refers to the first three scenarios which contain either A or C (or both).

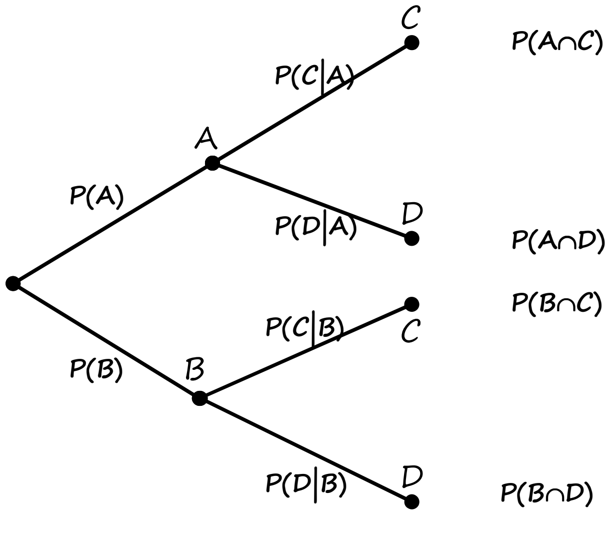
NOTICE

In the tree diagram above, the value **0.3** of the branch AC is in fact the conditional probability

$P(C|A)$ = Probability to obtain C, given that the first letter is A

In general, in a tree diagram

- the branches of the 1st column contain simple probabilities of the form $P(X)$
- the branches of the 2nd column contain conditional probabilities of the form $P(Y|X)$
- the results in the last column are combined probabilities of the form $P(X\cap Y)$



We may have more complicated tree diagrams, with more branches per level, more levels, etc.

EXAMPLE 1.

We throw a die.

If we get 1 we stop.

If we get 2,3,4 or 5 we toss a coin.

If we get 6 we toss two coins.

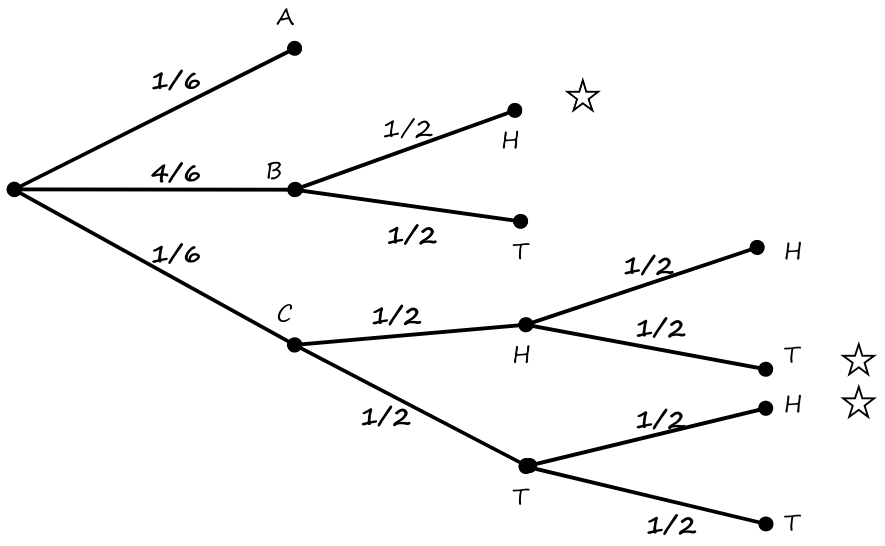
Find the probability that only one head is obtained.

Solution.

For our convenience, we denote the results of the die by

$$A=\{1\}, \qquad B=\{2,3,4,5\}, \qquad C=\{6\}$$

We construct the following tree diagram:



There are finally 7 scenarios (seven paths).

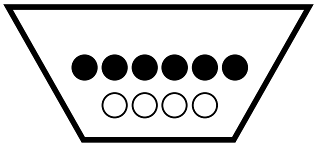
In 3 of them we have exactly one HEAD (we mark them by ☆)

We add the corresponding results:

$$\text{P(only one HEAD)} = \frac{4}{6}\cdot\frac{1}{2}+\frac{1}{6}\cdot\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{6}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{4}{12}+\frac{1}{24}+\frac{1}{24}=\frac{5}{12}$$

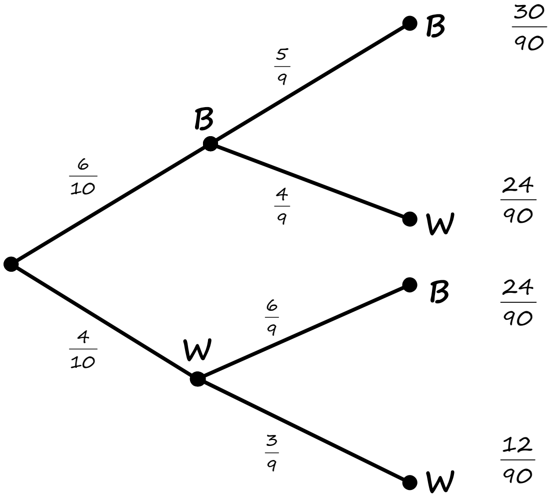
♦ A TYPICAL EXAMPLE: COLORED BALLS IN A BOX

A box contains 10 balls: 6 BLACK and 4 WHITE:



10 balls

We select two balls, one after the other. All possible outcomes are clearly shown on the following tree diagram



$$P(\text{both balls are BLACK}) = \frac{6}{10}\cdot\frac{5}{9} = \frac{30}{90} = \frac{1}{3}$$

$$P(\text{only one ball is BLACK}) = \left(\frac{6}{10}\cdot\frac{4}{9}\right)\times 2 = \frac{24}{90}\times 2 = \frac{8}{15}$$

$$P(\text{balls of same color}) = \frac{6}{10}\cdot\frac{5}{9} + \frac{4}{10}\cdot\frac{3}{9} = \frac{42}{90} = \frac{7}{15}$$

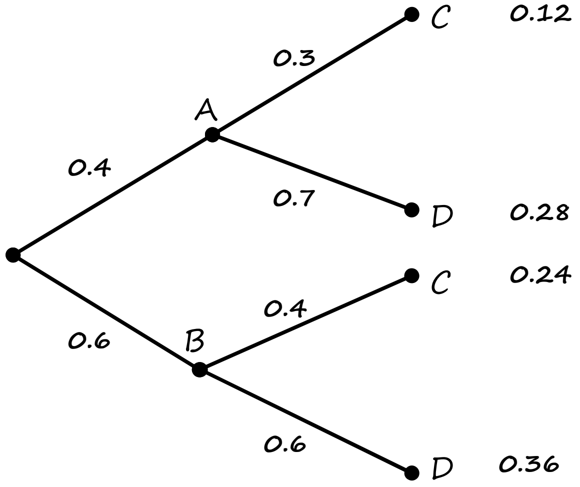
If we select 3 balls, we may follow the same rationale and answer directly without drawing a tree diagram. Thus,

$$P(\text{all three balls are BLACK}) = \frac{6}{10}\cdot\frac{5}{9}\cdot\frac{4}{8} = \frac{1}{6}$$

$$P(\text{only one ball is BLACK}) = \left(\frac{6}{10}\cdot\frac{4}{9}\cdot\frac{3}{8}\right)\times 3 = \frac{3}{10}$$

♦ THE "REVERSE GIVEN"

Consider again the tree diagram of PROBLEM 2



We said that $P(C|A) = 0.3$ is shown on the tree (on the branch AC).

What about $P(A|C)$?

Notice the "reverse" chronological order:

given that the final result is C,
what is the probability that the first result was A?

This result is not shown on the tree diagram; it is estimated as follows

$$P(A|C) = \frac{0.12}{0.12+0.24} \qquad \frac{\leftarrow \text{combination AC}}{\leftarrow \text{given C}}$$

Actually, it is the formula $P(A|C) = \dfrac{P(A \cap C)}{P(C)}$

Therefore,

$$P(A|C) = \frac{0.12}{0.36} \approx 0.33 \qquad P(B|C) = \frac{0.24}{0.36} \approx 0.67$$

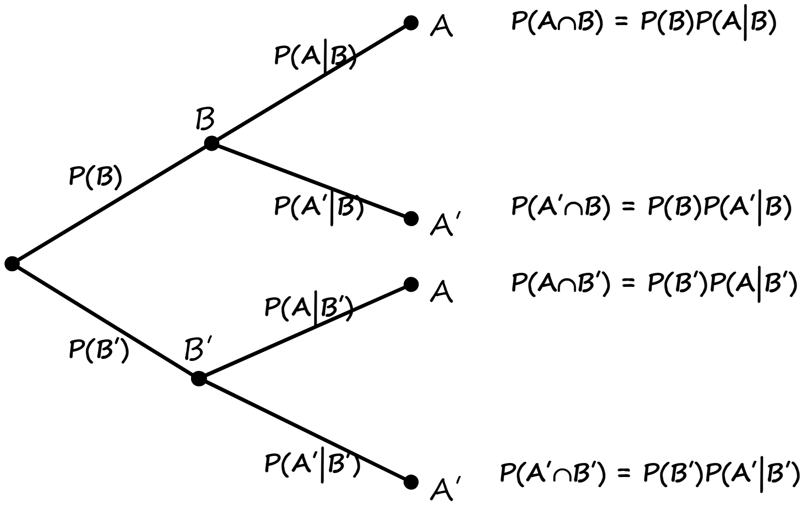
$$P(A|D) = \frac{0.28}{0.64} \approx 0.44 \qquad P(B|D) = \frac{0.36}{0.64} \approx 0.56$$

♦ BAYES' THEOREM **(only for HL)**

The "reverse given" result above is formally known as **Bayes' Theorem.**

The formula of this theorem can be easily obtained by a tree diagram.

The events in the first column are called B and B′
The events in the second column are called A and A′



Bayes' formula estimates the conditional probability P(B|A)

$$P(B \mid A)=\frac{P(B)P(A \mid B)}{P(B)P(A \mid B)+P(B')P(A \mid B')}$$

Assume now that the first level of the tree diagram consists of 3 disjoint events B_1,B_2,B_3. Bayes' theorem for $P(B_1|A)$ takes the form

$$P(B_1 \mid A)=\frac{P(B_1)P(A \mid B_1)}{P(B_1)P(A \mid B_1)+P(B_2)P(A \mid B_2)+P(B_3)P(A \mid B_3)}$$

Similar formulas apply for $P(B_2|A)$ and $P(B_3|A)$.

EXAMPLE 2.

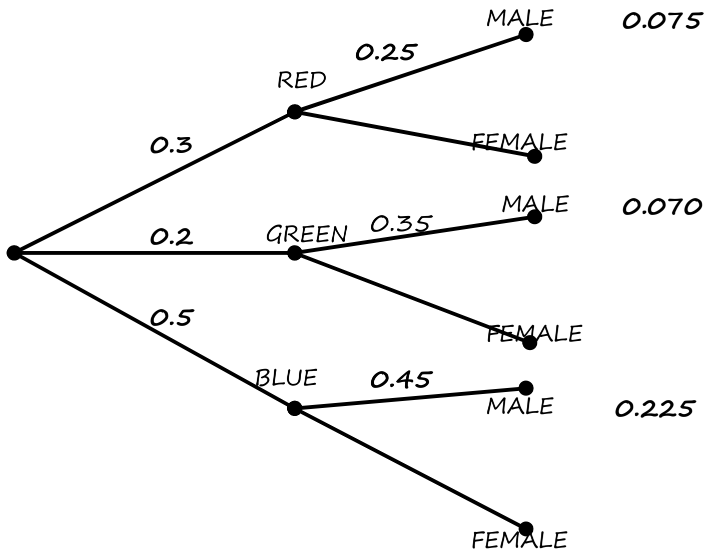
In a private school party, 30% of the students wear RED suits, 20% wear GREEN suits and 50% wear BLUE suits. 25% of the RED students, 35% of the GREEN students and 45% of the BLUE students are MALE. Find the probability that a MALE student wears GREEN suit, that is

$$P(GREEN|MALE).$$

Solution.

Instead of applying the Bayes' formula we will construct a tree diagram to obtain the "inverse given" probability.

Notice that we do not complete all the probabilities on the tree diagram, but only the "necessary" ones.



Therefore,

$$P(GREEN|MALE) = \frac{0.070}{0.075+0.070+0.225} = \frac{0.07}{0.37} \approx 0.189$$

In other words, 18.9% of the MALE students wear GREEN suits.

4.9 DISCRETE DISTRIBUTIONS IN GENERAL

Roughly speaking, **a random variable** X takes on some values in a given domain at random!!! It may be

Discrete	OR	**Continuous**
e.g. $X \in \{10,20,30\}$		e.g. $X \in [10,20]$
$X \in \{0,1,2,3,\ldots\}$		$X \in R$

A discrete variable takes on values in a finite or numerable set, while a continuous variable takes on values in some interval(s).

In this paragraph we only deal with discrete random variables.

♦ DISCRETE RANDOM VARIABLE

Let X be a variable which takes on the values

$$10, \quad 20, \quad 30$$

with probabilities

$$0.2, \quad 0.3, \quad 0.5$$

respectively. We often use a table

x	10	20	30
P(X=x)	0.2	0.3	0.5

Clearly

(i) all the probabilities are non-negative numbers; and

(ii) their sum is always 1.

Then we say that X is **a discrete random variable**.

To express that the probability that X=10 is 0.2 we write

$$P(X=10) = 0.2$$

Similarly, $P(X=20) = 0.3$ and $P(X=30) = 0.5$.

In general, for **a discrete random variable** X with

x	x_1 x_2 x_3 ...
P(X=x)	p_1 p_2 p_3 ...

it holds

(i) $p_i \geq 0$, for all *i*

(ii) $\sum p_i = 1$, i.e $p_1 + p_2 + p_3 + \cdots = 1$

We write

$P(X=x_1) = p_1$, $P(X=x_2) = p_2$, and so on.

(We also say that **a probability function** *p*: $x_i \mapsto y_i$ is defined).

♦ THE EXPECTED VALUE μ=E(X)

The **mean μ** or otherwise the **expected value E(X)** is defined by

$$E(X) = \sum x_i p_i = x_1p_1 + x_2p_2 + x_3p_3 + \cdots$$

For our example

x	10	20	30
P(X=x)	0.2	0.3	0.5

the expected value (otherwise the mean) is

$$E(X) = 10\times0.2 + 20\times0.3 + 30\times0.5 = 23$$

NOTICE: Explanation for μ=E(X)

In fact the mean here is not different than the mean in statistics

Consider the following ten numbers

10, 10, 20, 20, 20, 30, 30, 30, 30, 30

The probabilities to select 10, 20 or 30 are as in the table above.

The mean in statistics is also

$$\mu = \frac{10\times2+20\times3+30\times5}{10} = 10\times\frac{2}{10} + 20\times\frac{3}{10} + 30\times\frac{5}{10} = 23$$

EXAMPLE 1

Consider

x	10	20	30
P(X=x)	a	b	0.5

Given that E(X)=23, find the values of a and b.

Solution.

We use two relations

$$a + b + 0.5 = 1 \quad \Rightarrow \quad a + b = 0.5$$

$$10a + 20b + 30\times 0.5 = 23 \quad \Rightarrow \quad 10a+20b=8$$

The solution of the system is $a = 0.2$ and $b = 0.3$

The probability distribution applies in many betting games:

EXAMPLE 2

Consider again the same table above. But now we select one of the numbers 10, 20, 30 at random.

If we select 10 we earn 6 points

If we select 20 we earn 1 point

If we select 30 we lose 2 points

What is the expected number of points in one game?

Solution.

We extend our table as follows

x	10	20	30
Profit	6 points	1 point	-2 points
Prob	0.2	0.3	0.5

We estimate the expected profit:

$$\text{Expected profit} = 6\times 0.2 + 1\times 0.3 - 2\times 0.5 = 0.5$$

That is, in each game we earn 0.5 points on average.

Explanation

In other words, if we play this game 10 times we expect to earn 5 points on average.

Indeed, if we play the game 10 times we expect to obtain

2 times the number 10, that is 2×6=12 points

3 times the number 20, that is 3×1=3 points

5 times the number 30, that is 5×(–2)=–10 points

In total, 12+3–10 = 5 points

EXAMPLE 3

We throw two dice.

If we obtain TWO SIXES	we earn 15€
If we obtain ONLY ONE SIX	we earn 1€
If we obtain NO SIX	we lose 1€

Find the expected profit in one game.

Solution.

Let us organize our data on a table

Result	TWO SIXES	ONE SIX	NO SIX
Profit	15€	1€	–1€
Prob	$\frac{1}{36}$	$\frac{10}{36}$	$\frac{25}{36}$

The expected amount earned per game is

$$\text{Expected profit} = 15\times\frac{1}{36}+1\times\frac{10}{36}-1\times\frac{25}{36}=0$$

This is a FAIR GAME! We expect neither to earn nor to lose!

Notice. If the first winning prize was not 15€ but 14€, the expected profit would be $-\frac{1}{36}$

In other words, if we play the game 36000 times (or otherwise bet 36000€) we expect to lose 1000€.

♦ MEDIAN-MODE

These measures, known from statistics, are defined analogously:

MODE = The value X=a of the highest probability

MEDIAN = The value X=m where the probability splits in two equal parts (0.5-0.5)

Look at the examples below

x	10	20	30
P(X=x)	0.4	0.3	0.3

MODE = 10

MEDIAN = 20

x	10	20	30
P(X=x)	0.2	0.3	0.5

MODE = 30

MEDIAN = 25 (why?)

♦ VARIANCE (**Only for HL**)

We define

$$Var(X) = E(X-\mu)^2$$

that is

$$Var(X) = (x_1-\mu)^2\times p_1 + (x_2-\mu)^2\times p_2 + (x_3-\mu)^2\times p_3 + \dots$$

An equivalent definition is

$$Var(X) = E(X^2)-\mu^2$$

where

$$E(X^2) = x_1{}^2\times p_1 + x_2{}^2\times p_2 + x_3{}^2\times p_3 + \dots$$

EXAMPLE 4

Consider again the probability distribution

x	10	20	30
P(X=x)	0.2	0.3	0.5

We have seen that μ=E(X)=23. Therefore,

$$Var(X) = (10-23)^2\times0.2+(20-23)^2\times0.3+(30-23)^2\times0.5 = 61$$

or

$$E(X^2) = 10^2\times0.2+20^2\times0.3+30^2\times0.5 = 590$$

$$Var(X)=590-23^2 = 61$$

4.10 BINOMIAL DISTRIBUTION – B(n,p)

It is the distribution of a discrete random variable X which takes on the values

0, 1, 2, 3, 4, … , n

with probability function

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0,1,2,3,\ldots,n$$

where n, p are two parameters. We will see that the binomial distribution describes a certain type of problems.

Notice: the formula **is not** in the syllabus; Results will be obtained directly by GDC. It is worth to mention though how it works!

- DESCRIPTION OF THE PROBLEM

We deal with a game (or any experiment) with two outcomes

SUCCESS with probability p
FAILURE (with probability 1-p)

We play the game n times. Our parameters are

n = number of trials
p = probability of success

while

X counts the number of (possible) successes

We say that X follows a binomial distribution and write X~B(n,p).

Since n is the number of trials, X can take on the values

0, 1, 2, 3, 4, …, n

The probabilities P(X=0), P(X=1), P(X=2), etc can be obtained by the GDC.

(and also by the formula mentioned in the introduction, but as we have said this formula is not in the syllabus).

♦ GDC

Our GDC (Casio) gives the results for a Binomial distribution

MENU – Statistics – DIST – BINOMIAL: We use **Bpd** or **Bcd**

For simplicity let us denote by

Bpd(x)	the probability of exactly x successes
Bcd(x_1 to x_2)	the probability from x_1 up to x_2 successes

The menu for both functions is

Data:	always **Variable**
Numtrial:	is the number of trials, i.e. **n**
p:	is the probability of success **p** (for each game)

Then for each value of x (or x_1 to x_2), EXE gives the result.

EXAMPLE 1

We toss a die 5 times. The success is to get a **six**. Then

$$n=5 \quad \text{and} \quad p=\frac{1}{6}$$

We may have 0, 1, 2, 3, 4 or 5 successes.

The probability distribution for X is given by (results in 4dp)

x	0	1	2	3	4	5
GDC	Bpd(0)	Bpd(1)	Bpd(2)	Bpd(3)	Bpd(4)	Bpd(5)
P(X=x)	0.4019	0.4019	0.1608	0.0322	0.0032	0.0001

We can also answer the following questions:

Find the probability of	Notation	GDC	Result
exactly 3 sixes	P(X=3)	**Bpd(3)**	0.0322
at most 3 sixes	P(X≤3)	**Bcd(0 to 3)**	0.9967
less than 3 sixes	P(X<3)	**Bcd(0 to 2)**	0.9645
more than 3 sixes	P(X>3)	**Bcd(4 to 5)**	0.0033
at least 3 sixes	P(X≥3)	**Bcd(3 to 5)**	0.0355

Remark for the formula (not in the syllabus but worth to know)

The probability to obtain

5 sixes in a row is	$(1/6)^5$
no six at all is	$(5/6)^5$
2 sixes and 3 no-sixes is	$\binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3$

$\binom{5}{2}$ (or 5C2) is the number of ways to have 2 sixes in 5 trials.

In general, the probability to obtain x sixes (and (5-x) no-sixes) is

$$\binom{5}{x}\left(\frac{1}{6}\right)^x\left(\frac{5}{6}\right)^{5-x}$$

In general, if we play n times a game with probability of success p

the probability P(X=x) is given by the formula

$$p(X=x)=\binom{n}{x}p^x(1-p)^x$$

According to the formula

x	0	1	2	3	4	5
P(X=x)	$\frac{3125}{6^5}$	$\frac{3125}{6^5}$	$\frac{1250}{6^5}$	$\frac{250}{6^5}$	$\frac{25}{6^5}$	$\frac{1}{6^5}$
	=0.4019	=0.4019	=0.1608	=0.0322	=0.0032	=0.0001

The table agrees with the results found by Bpd(x) above.

♦ EXPECTED VALUE AND VARIANCE OF X

They are given by the formulae

$$\mathbf{E(X) = np} \qquad \mathbf{Var(X) = np(1-p)}$$

For our example above

$$E(X) = 5\frac{1}{6} = \frac{5}{6} \qquad \text{and} \qquad Var(X) = 5\frac{1}{6}\,\frac{5}{6} = \frac{25}{36}$$

Notice (only for HL)

Since you know E(X) and Var(X), you also know $E(X^2)$. Indeed,

$$Var(X) = E(X^2) - E(X)^2$$
$$\Rightarrow \quad E(X^2) = Var(X) + E(X)^2$$
$$\Rightarrow \quad E(X^2) = np(1-p) + (np)^2$$

EXAMPLE 2

A box contains 5 balls, 1 BLACK and 4 WHITE. We win if we select a BLACK ball. We play this game 10 times.

Find

(a) The probability to win exactly 4 times

(b) The probability to win at most 4 times

(c) The probability to win at least once

(d) The expected number of winning games.

(e) The variance of the number of winning games.

Solution

The variable

X = number of winning games

follows a binomial distribution with n=10 and $p=\frac{1}{5}=0.2$

[we may also write X~B(10,0.2)]

(a) The probability to win exactly 4 times is **Bpd(4)=0.088**

[indeed, $P(X=4) = \binom{10}{4}(0.2)^4(0.8)^6 = 0.088$]

(b) The probability to win at most 4 times is **Bcd(0 to 4)=0.967**

[in fact $P(X\le 4) = P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4)$]

(c) The probability to win al least once is **Bcd (1 to 10) = 0.893**

[in fact $P(X\ge 1) = 1-P(X=0) = 1-0.107 = 0.893$]

(d) The expected number is **E(X)=np=10×0.2 = 2**

(e) The variance is **Var(X) = np(1-p) = 10×0.2×0.8 = 1.6**

EXAMPLE 3

Let p=0.2 and n unknown. It is given that P(X=1) = 0.268. Find n.

Solution

We know that n must be an integer.

By trial and error on Numtrial we can see that Bpd(10)=0.268

Hence n=10.

- MODE (**mainly for HL**)

  We first check the expected number

  - If the expected number is in decimal form,

    say n=20, $p=\frac{1}{6}$, so that $E(X)=\frac{20}{6}\cong 3.3$

    we check the nearest integer values 3 and 4

    P(X=3) = 0.237

    P(X=4) = 0.202

    Hence the **mode is 3** (it has the highest probability)

  - If the expected number is a whole number,

    say n=60, $p=\frac{1}{6}$, so that $E(X)=\frac{60}{6}=10$

    we check the neighboring integer values 9, 10, 11

    P(X=9) = 0.134

    P(X=10) = 0.137

    P(X=11) = 0.126

  Hence the **mode is 10**.

Notice In some cases we may have two modes:

For n=5 and $p=\frac{1}{6}$, it is $E(X)=\frac{5}{6}=0.833$. We check

P(X=0) = 0.4019

P(X=1) = 0.4019

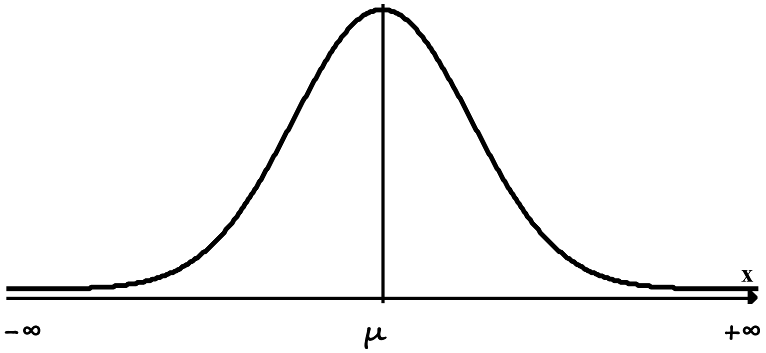
Hence there are two modes, 0 and 1.

4.11 NORMAL DISTRIBUTION – N(μ,σ^2)

It is the distribution of a **continuous random variable** X with values form -∞ to +∞. The parameters of this distribution are

μ = mean
σ = standard deviation.

The "behavior" of the probability is described by a function which looks like



Roughly speaking, there is a highly likely mean value μ and all the other values of X spread out symmetrically about the mean. As we move away from the mean (either to the left or to the right of the mean) the probability decreases dramatically!

We say that X follows a normal distribution with mean μ and standard deviation σ (or variance σ^2) and we write X~N(μ,σ^2).

♦ DESCRIPTION OF THE PROBLEM IN GENERAL

It is the most "popular" distribution in nature. Random variables which depend on many factors follow this distribution, for example

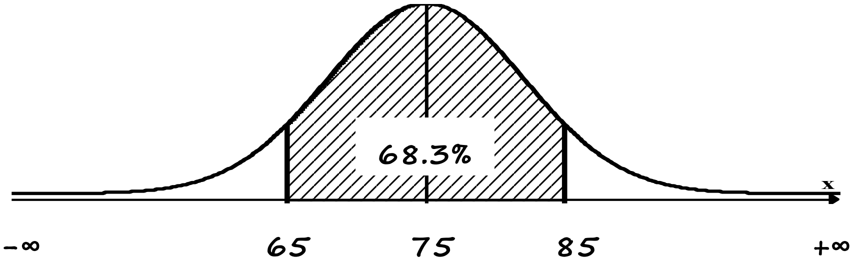
- Weight of people
- Height of people
- Time spent in a super market
- Weight of a pack of coffee labeled 500 g.

For example, suppose that for a Greek man

mean weight: **μ=75kg** st.dev. of the weights: **σ=10kg**

It is estimated[4] that

Percentage of the population	**ranges between**	
	in general	**for our problem**
about 68% of the population	$\mu-\sigma$ and $\mu+\sigma$	[65,85]
about 95% of the population	$\mu-2\sigma$ and $\mu+2\sigma$	[55,95]
about 99.7% of the population	$\mu-3\sigma$ and $\mu+3\sigma$	[45,105]



NOTICE

- The whole area under the curve is 1 (i.e. 100%). The area before the mean as well as the area after the mean is 0.5 (i.e. 50%)
- Theoretically, the distribution of X ranges between -∞ to +∞. In practice, we may assume that almost the whole population (in fact 99,7%) ranges between $\mu-3\sigma$ and $\mu+3\sigma$.
- The standard deviation σ indicates the spread of the population. For example, assume that

Greeks:	μ=75 kg	σ=10 kg
Italians:	μ=75 kg	σ=8 kg

  This implies that both populations have the same mean but Italians are closer to the mean than Greeks. In other words, almost the whole population is between $\mu\pm3\sigma$, namely

75±30 i.e. 45-105 kg	for Greeks
75±24 i.e. 51-99 kg	for Italians

[4] We will explain in a while how we get the following estimations.

We will distinguish three types of problems.
In all these problems we use the GDC in order to find the results.
For Casio fx

MENU – STAT – DIST – NORM: We use **Ncd** or **InvN**
Data: always use **Variable**

In general, **Ncd** is used when we ask for a probability
InvN is used when we know the probability

♦ PROBLEM 1: FIND PROBABILITY (μ,σ known, we use **Ncd**)

Consider again the example where

X = the weight of a Greek man

with μ=75 kg and σ=10 kg.

Find the probability that a Greek man weighs

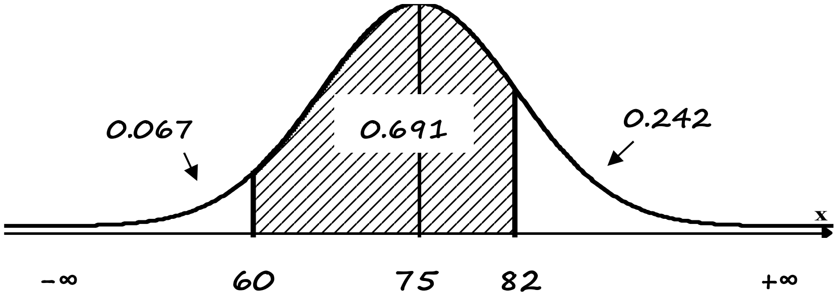
(a) between 60 and 82 kg [that is $P(60 \le X \le 82)$]
(b) more than 82 kg [that is $P(X \ge 82)$]
(c) less than 60 kg [that is $P(X \le 60)$]

Solution

We use Ncd in the GDC. We set **σ=10, μ=75**

Question	Ncd		Press EXE
(a)	Lower 60	Upper 82	0.691
(b)	Lower 82	Upper 999999…	0.242
(c)	Lower -999999	Upper 60	0.067

Let us represent the information of this problem by a diagram:



NOTICE

- For $P(60 \le X \le 82)$ the GDC gives $p=0.691$ (question (a))

  Below this result some extra information is given:

  $$z{:}Low = -1.5 \quad z{:}Up = 0.7$$

  This means that

  the lower bound is 1.5 st. devs below μ: $75-1.5\times10=60$

  the upper bound is 0.7 st. devs above μ: $75+0.7\times10=82$

  We will refer to those values z later on; they are known as standardized values.

- The probability that the weight is 1 standard deviation away from the mean, that is between 65 and 85 kg is

  $P(65 \le X \le 85) \cong 0.683$, (68.3%) as we said in the introduction.

  Notice that $z{:}Low = -1$ $z{:}Up = 1$

- The probabilities above refer to a selection of one person only.

  For example,

  P(a person is between 60 and 82 kg) = 0.691,

  P(a person is not between 60 and 82 kg) =1-0.691= 0.309

  If we select two people,

  P(both between 60 and 82 kg) = $(0.691)^2$

  P(none between 60 and 82 kg) = $(0.309)^2$

  P(only one between 60 and 82 kg) = 2× (0.691)×(0.309)

- A question may combine the normal and binomial distributions: Suppose that we select 10 people. What is the probability that exactly three of them are between 60 and 82 kg?

  The normal distribution gives $p=0.691$.

  A binomial distribution (of a new variable Y) is defined with

  $$n=10 \text{ and } p=0.691.$$

  Hence, for $Y \sim B(10,0.691)$

  $$P(Y=3) = 0.0106$$

♦ PROBLEM 2: PROBABILITY IS GIVEN (μ,σ known, we use **InvN**)

Again, let μ=75 kg and σ=10 kg for the variable

X = the weight of a Greek man

The probability that somebody weighs less than a is 0.067. That is

$P(X \le a) = 0.067$

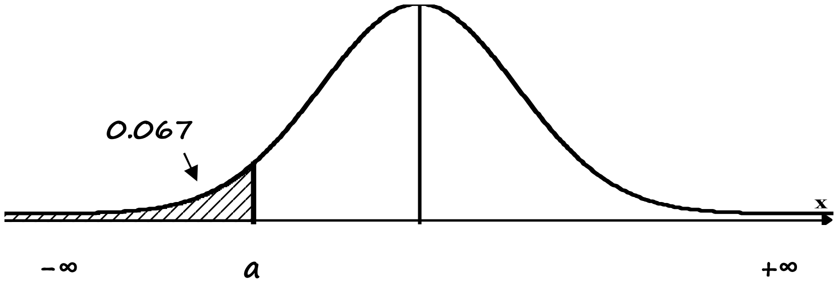
Find a.

(An alternative way to set the same question is

6.7% of the Greek men weigh less than a. Find a)

Solution

Let us represent this information in a diagram



We use **InvN**. We set the parameters **σ=10, μ=75**. Then

Tail: **Left** (it is the area before a)

Area: 0.067

Press **EXE** and obtain **a=60** kg.

Notice for the tail

Tail: Left	If we know the area before some value
Tail: Right	If we know the area after some value

Hence, the answer above may obtained by using right tail.

The area before a is 0.067, so the area after a is 0.933

In other words $P(X \ge a) = 0.933$. Then

Tail: **Right** (it is the area after a)

Area: 0.933

Press **EXE** and obtain **a=60** kg.

EXAMPLE 1

The mass of packs of a certain type of coffee is normally distributed with a mean of 500 g and standard deviation of 15 g.

(a) Find the probability that a pack weighs more than 520 g

(b) The lightest 4% of the packs weigh less than a, while the heaviest 4% of the packs weigh more than b. Find a and b.

The packs in question (b) are rejected from the market.

(c) In a daily production of 1600 packs how many of them are expected to be rejected?

(d) We select 2 packs. Find the probability that both are rejected.

(e) We select 5 packs. Find the probability that al least one is rejected.

(f) Find Q_1 and Q_3, the lower and upper quartiles of the weights

Solution

(a) We use Ncd

$P(X\geq520) \cong 0.091$

(b) We use InvN

$P(X\leq a)=0.04$, hence $a \cong 474$ g

$P(X\geq b)=0.04$, hence $b \cong 526$ g

(c) 8% is rejected, thus 1600x0.08=128 packs

(d) $(0.08)^2=0.0064$

(e) Binomial with n=5 and p=0.08 (i.e. 8%)

$P(X\geq1) \cong 0.341$

(f) In fact, it looks like question (b). We know that the "area" before Q_1 is 0.25, while before Q_3 is 0.75. We use InvN

$P(X\leq Q_1)=0.25 \Rightarrow Q_1 \cong 490$ g

$P(X\leq Q_3)=0.75 \Rightarrow Q_3 \cong 510$ g

Particularly for this result we can use **Tail:Central**, **Area =0.5**

For Problem 3 we need first a procedure called standardization.

♦ STANDARDISATION – NORMAL DISTRIBUTION N(0,1)

Consider the random variable Z which follows Normal distribution with

$$\mu=0 \text{ and } \sigma=1$$

This is **the standardised normal distribution**: $Z \sim N(0,1)$

Any variable X that follows normal distribution can be transformed into the standardised normal variable Z by using the formula

$$Z=\frac{X-\mu}{\sigma}$$

For our example:

X follows $N(\mu,\sigma^2)$ with $\mu=75$ and $\sigma=10$,

$Z=\frac{X-75}{10}$ follows N(0,1).

We also say that,

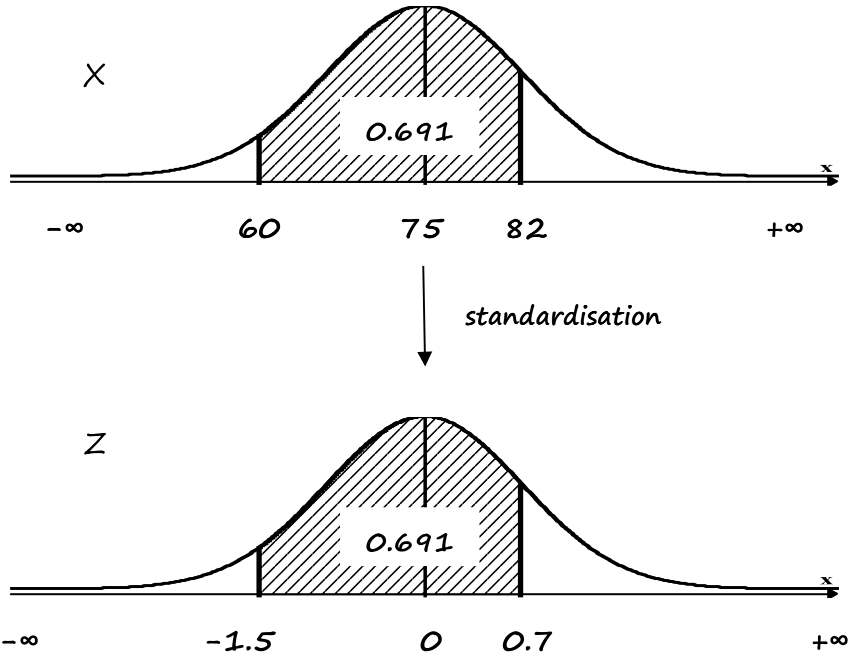
the standardised value of x=60 is $z_{60}=\frac{60-75}{10}=-1.5$

the standardised value of x=82 is $z_{82}=\frac{82-75}{10}=0.7$

Check by your GDC that

$P(60\le X\le 82)=0.691$ and $P(-1.5\le Z\le 0.7) = 0.691$

In diagrams:



- PROBLEM 3: FIND μ OR σ (we use **Standardisation** and **InvN**)

A random variable X follows normal distribution with

$$\mu=150 \quad \text{and} \quad \sigma \text{ unknown.}$$

It is given that 25% is less than 140, that is $P(X\leq 140)=0.25$

(In other words, the probability that X is less than 140 is 0.25)

Find σ.

Solution

We use the formula

$$Z=\frac{X-\mu}{\sigma}.$$

For X=140, Z can be obtained by the GDC and μ=150

The standardized value Z_{140} is obtained by the GDC, using **InvN:**

Tail: **Left** (we know the area before 140)

Area: 0.25

σ=1

μ=0

Press **EXE** and obtain $Z_{140}=-0.6745$

Then

$$-0.6745=\frac{140-150}{\sigma}$$
$$\Rightarrow \sigma=\frac{-10}{-0.6745}$$
$$\Rightarrow \sigma\cong 14.83$$

A similar process can be followed if μ is unknown.

For the same problem above suppose that

$$\sigma=14.83 \quad \text{but} \quad \mu \text{ unknown.}$$

The standardized value Z_{140} is exactly the same: $Z_{140}=-0.6745$

Then

$$-0.6745=\frac{140-\mu}{14.83}$$
$$\Rightarrow 140-\mu \cong -10$$
$$\Rightarrow \mu\cong 150$$

Let us see an example where both μ and σ are unknown.

EXAMPLE 2

For a random variable X we know that

35% is less than 60

25% is more than 90

That is

$$P(X \le 60)=0.35, \quad P(X \ge 90)=0.25.$$

Find μ and σ.

Solution

(The diagram is not necessary - sometimes it helps)

The standardized values Z_{60} and Z_{90} can be obtained by the GDC:

Tail: Left, Area: 0.35, σ=1, μ=0 gives $Z_{60} = -0.385$

Tail: Right, Area: 0.25, σ=1, μ=0 gives $Z_{90} = 0.674$

Then

$$-0.385 = \frac{60-\mu}{\sigma} \quad \text{and} \quad 0.674 = \frac{90-\mu}{\sigma}$$

We obtain the system

$$\mu - 0.385\sigma = 60$$

$$\mu + 0.674\sigma = 90.$$

The solution of the system is μ=70.9 and σ= 28.3

Notice (only for HL)

Since you know E(X) and Var(X), you also know $E(X^2)$. Indeed,

$$\begin{aligned} & Var(X) = E(X^2) - E(X)^2 \\ \Rightarrow \quad & E(X^2) = Var(X) + E(X)^2 \\ \Rightarrow \quad & E(X^2) = \sigma^2 + \mu^2 \end{aligned}$$

ONLY FOR

HL

4.12 CONTINUOUS DISTRIBUTIONS IN GENERAL (for HL)

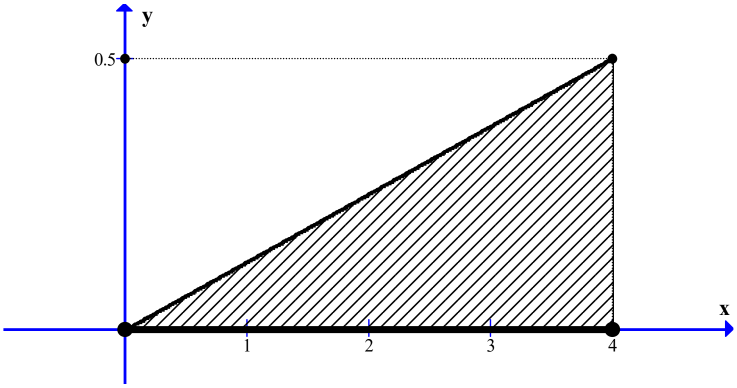
Let X be a variable which takes on values in the interval

$$[0,4]$$

Suppose also that the probability for the value of X is not "uniformly" distributed throughout this interval but it is more likely that X obtains values near 4.

In such a case we have a continuous function which describes the behavior of the probability. Assume that this function is

$$f(x)=\frac{x}{8}, \quad 0\le x\le 4$$



That is, the probability increases as we move towards 4.

The function is not accidental! Notice that

(i) $f(x)\ge 0$

(ii) the area of the triangle under the graph is 1

The probability that X is between 0 and 2 is given by the corresponding area under the curve, that is 0.25 (25% of the total area). We write

$$P(0\le X\le 2) = 0.25$$

Similarly

$$P(2\le X\le 4) = 0.75$$

Then we say that X is **a continuous random variable**. The function f(x) is called **probability density function (pdf)**.

For a continuous random variable we measure only the probability of an interval, not of a fixed value; we agree that the probability that X takes on a particular value a is 0, that is

$$P(X=a) = 0$$

In general, for **a continuous random variable** X with

probability density function (or **pdf**) $f(x)$

it holds

(i) $f(x) \geq 0$, i.e. the function is non-negative

(ii) $\int_{-\infty}^{+\infty} f(x)dx = 1$, i.e the total area under the curve is 1

while the probability that X takes values between a and b is

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Notice that

$$P(a \leq X \leq b) \quad \text{and} \quad P(a<X<b)$$

are exactly the same as the probability that X takes a particular value, say P(X=a) is zero!

For our example,

$$f(x) = \frac{x}{8}, \quad 0 \leq x \leq 4$$

it holds

$$\int_{-\infty}^{+\infty} f(x)dx = \int_0^4 \frac{x}{8}dx = \left[\frac{x^2}{16}\right]_0^4 = 1$$

and

$$P(0 \leq X \leq 2) = \int_0^2 \frac{x}{8}dx = \left[\frac{x^2}{16}\right]_0^2 = \frac{4}{16} = 0.25$$

$$P(2 \leq X \leq 4) = 1 - 0.25 = 0.75$$

Notice also that for particular values of X, the probability is 0. For example

$$P(X=2) = 0, \qquad P(X=3.7) = 0$$

- THE EXPECTED VALUE μ=E(X)

The **mean μ**, or otherwise the **expected value E(X)** is defined by

$$\mu=E(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

For our example

$$\mu=E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_{0}^{4} \frac{x^2}{8}dx = \left[\frac{x^3}{24}\right]_0^4 = \frac{8}{3} \quad (\cong 2.67)$$

- THE VARIANCE Var(X)

It is defined by

$$Var(X) = E(X-\mu)^2$$

that is

$$Var(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx$$

An equivalent (and more practical) definition is

$$Var(X) = E(X^2)-\mu^2$$

where

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)dx$$

For our example,

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x)dx = \int_{0}^{4} \frac{x^3}{8}dx = \left[\frac{x^4}{32}\right]_0^4 = 8$$

thus

$$Var(X) = E(X^2)-\mu^2 = 8 - \frac{8}{3} = \frac{16}{3}$$

Notice:

The initial definition gives

$$Var(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx = \int_{0}^{4} (x-\frac{8}{3})^2 \frac{x}{8}dx$$

Which is more complicated to calculate.

NOTICE. Compare the definitions for discrete and continuous X

X DISCRETE	X CONTINUOUS
μ=E(X) = $\sum x_i p_i$	μ=E(X) = $\int_{-\infty}^{+\infty} xf(x)dx$
E(X²) = $\sum x_i^2 p_i$	E(X²) = $\int_{-\infty}^{+\infty} x^2 f(x)dx$
Var(X) = E(X²)-μ^2	

- MODE

It is the value of x where f(x) has its maximum.

For our example,

$$\text{MODE} = 4$$

- MEDIAN

It is the value of m where P(X≤m)=0.5

In practice, we find m by solving

$$\int_{-\infty}^{m} f(x)dx = 0.5$$

For our example,

$$\int_0^m \frac{x}{8}dx = 0.5 \Leftrightarrow \left[\frac{x^2}{16}\right]_0^m = 0.5 \Leftrightarrow \frac{m^2}{16} = 0.5 \Leftrightarrow m^2 = 8 \Leftrightarrow m = \sqrt{8}$$

- QUARTILES

The **lower quartile Q_1** and the **upper quartile Q_3** are defined by

$$P(X\le Q_1)=0.25 \qquad P(X\le Q_3)=0.75$$

Namely,

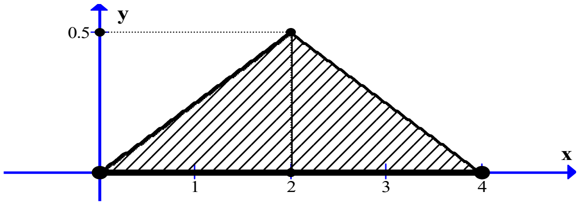
$$\int_{-\infty}^{Q_1} f(x)dx = 0.25 \quad \text{and} \quad \int_{-\infty}^{Q_3} f(x)dx = 0.75$$

For our example, Q_1 = 2 and Q_3 = $2\sqrt{3}$ (why?)

Look at the following step function.

EXAMPLE 1

Let X be a continuous random variable in [0,4] with pdf



$$f(x)=\begin{cases} \frac{x}{4}, & 0\le x\le 2 \\ 1-\frac{x}{4}, & 2\le x\le 4 \end{cases}$$

- Let us confirm that f(x) is a pdf:

$$\int_{-\infty}^{+\infty} f(x)dx \;=\; \int_0^2 \frac{x}{4}dx+\int_2^4 (1-\frac{x}{4})dx \;=\; \ldots \;=\; \frac{1}{2}+\frac{1}{2}=1$$

[in fact, it would be easier to find the area from the graph!].

- The expected value is

$$\mu{=}E(X) \;=\; \int_{-\infty}^{+\infty} xf(x)dx \;=\; \int_0^2 \frac{x^2}{4}dx+\int_2^4 (x-\frac{x^2}{4})dx \;=\; \ldots \;=\; 2$$

[in fact, it is obvious by the symmetry of the graph that μ=2]

- For the variance, we find first

$$E(X^2) \;=\; \int_{-\infty}^{+\infty} x^2f(x)dx \;=\; \int_0^2 \frac{x^3}{4}dx+\int_2^4 (x^2-\frac{x^3}{4})dx \;=\; \ldots \;=\; 1+\frac{11}{3}=\frac{14}{3}$$

Then $\mathrm{Var}(X) = \frac{14}{3}-2^2=\frac{2}{3}$

- Since $\int_0^2 \frac{x}{4}dx = 0.5$, the median is 2.

Notice. Let $f(x)=\begin{cases} f_1(x), & a\le x\le b \\ f_2(x), & b\le x\le c \end{cases}$. We check $\int_a^b f_1(x)dx$ =A

If A > 0.5, the median is between a and b, we solve $\int_a^{median} f_1(x)dx=0.5$

If A < 0.5, the median is between b and c, we solve $\int_{median}^{c} f_2(x)dx=0.5$

4.13 COUNTING – PERMUTATIONS – COMBINATIONS (for HL)

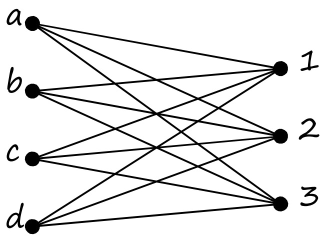
♦ MULTIPLICATION PRINCIPLE – BOX TECHNIQUE

Suppose that

Task A has 4 possible outcomes: a,b,c,d

Task B has 3 possible outcomes: 1,2,3

The outcomes for A and B can be combined in the following ways



a1	b1	c1	d1
a2	b2	c2	d2
a3	b3	c3	d3

The number of combinations is 12

This situation can be represented by using the BOX-TECHNIQUE:

A	B
4	3

There are 4 ways to complete box A, 3 ways to complete box B. The result is given by 4×3 = 12

In general,

A	B
m	n

m choices for A, **n** choices for B ⇒ **mxn** choices altogether

EXAMPLE 1

Two dice are tossed. How many results are there?

Each die has 6 outcomes. Hence,

$$\boxed{6}\ \boxed{6}$$

6×6 = **36** results altogether

In this example it would be easy to write down all possible combinations of dice and realize that their number is 36. However, it is not always practical to count all the possible outcomes.

EXAMPLE 2

The Latin alphabet has 26 letters. How many combinations of two letters are there if

a) repetition of letters is allowed

b) no repetition is allowed (i.e. different letters)

a) there are 26 choices for each box

$$\boxed{26}\ \boxed{26}$$

26×26=**676** combinations altogether

b) there are 26 choices for the first box. Once we complete the first box there are 25 choices for the second box

$$\boxed{26}\ \boxed{25}$$

26×25=**650** combinations altogether

♦ WORDS OF r LETTERS

Consider an alphabet of n letters. Then

$$\boxed{\text{number of words of } r \text{ letters} = n^r}$$

Indeed, let us place r boxes in a row.

$$\boxed{n}\ \boxed{n}\ \boxed{n}\ \boxed{\ldots}\ \boxed{n}$$

There are n choices for each box. Therefore, the total number of words is

$$\underbrace{n\cdot n\cdot n\cdots n}_{r\ \text{times}} = n^r$$

EXAMPLE 3

Find the total number of words with 4 letters if our alphabet is

a) the alphabet of the 26 Lattin letters
b) only the letters A,B,C,D,E
c) the ten digits 0,1,2,3,4,5,6,7,8,9
d) only the digits 0,1 (known as binary alphabet)

Solution

The total number of words for each case is

a) $26^4 = 456976$
b) $5^4 = 625$
c) $10^4 = 10000$
d) $2^4 = 16$

Instead of remembering the formula n^r it is sometimes more convenient to use directly the box-technique. For example, in d) above we have 4 boxes and 2 choices, 0 or 1, for each box:

$$\boxed{2}\ \boxed{2}\ \boxed{2}\ \boxed{2} \longrightarrow 2^4 = 16$$

EXAMPLE 4

A password has the form XXYYY where

X is one of the 26 Latin letters

Y is one of the digits 0,1,2,3,4,5,6,7,8,9

The total number of possible passwords is

$$\boxed{26}\ \boxed{26}\ \boxed{10}\ \boxed{10}\ \boxed{10}$$

$$26x26x10x10x10 = 26^2 10^3 = 676000$$

♦ FACTORIAL: n!

A new symbol is introduced.

We define

$$n! = 1\cdot2\cdot3\cdots n$$

which is read "n factorial".

For example

$$1! = 1$$
$$2! = 1\cdot2 = 2$$
$$3! = 1\cdot2\cdot3 = 6$$
$$4! = 1\cdot2\cdot3\cdot4 = 24$$
$$5! = 1\cdot2\cdot3\cdot4\cdot5 = 120$$

We also agree that $0! = 1$

(it looks peculiar, I know! Please accept it!)

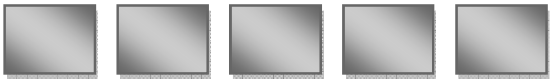
NOTICE: GDC can be used for the calculation of x!

Select RUN in the MENU: OPTN – PROB – x!

♦ REARRANGEMENTS OF n OBJECTS

Consider 5 objects, say A,B,C,D,E. How many ways are there to arrange them in order?

We use the box technique!



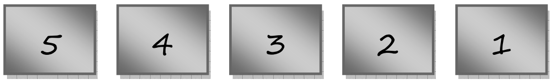
We have **5 choices** for the 1st box: A or B or C or D or E

Once it is completed,

we have **4 choices** for the 2nd box,

and so on.

Therefore,



there are $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5!$ possible arrangements

In general

n objects can be arranged in n! ways

EXAMPLE 5

Three people, **A**lex, **B**ill, **C**hris must sit in three chairs in a row! There are 3!=6 ways for them to be arranged. Indeed,

ABC ACB BAC BCA CAB CBA

♦ COMBINATIONS AND PERMUTATIONS

Consider n objects. How many ways are there to select r out of them? Two situations arise

COMBINATIONS	**PERMUTATIONS**
We do not mind about the order (The r objects are seen as a group)	We mind about the order (AB is different than BA)
nCr	**nPr**

The number of COMBINATIONS is denoted by **nCr** ("n choose r ")
The number of PERMUTATIONS is denoted by **nPr**

Let us explain the two situations by a simple example

EXAMPLE 6

Consider n=5 objects, A,B,C,D,E. We select r =2 out of them.

There are 10 possible COMBINATIONS

AB			
AC	BC		
AD	BD	CD	
AE	BE	CE	DE

There are 20 possible PERMUTATIONS (we mind about the order)

AB	BA	CA	DA	EA
AC	BC	CB	DB	EB
AD	BD	CD	DC	EC
AE	BE	CE	DE	ED

That is why **5C2 = 10** and **5P2 = 20**

As the number of objects increases we realize that it is not very convenient to write down the possible outcomes.

We can directly apply the formulas

$$nCr = \binom{n}{r} = \frac{n!}{r!(n-r)!} \qquad nPr = \frac{n!}{(n-r)!}$$

For example $5C2=\frac{5!}{2!3!}=10$ while $5P2=\frac{5!}{3!}=20$

NOTICE

$\binom{n}{1}=n$ Indeed, there are **n ways** to choose 1 out of n

[the 1st,or the 2nd, or the 3rd, or …, or the nth]

$\binom{n}{n}=1$ Indeed, there is only **1 way** to choose n out of n

[that is to choose all the n objects]

$\binom{n}{0}=1$ Indeed, there is only **1 way** to choose 0 out of n

[that is to choose nothing!]

Notice also that

$$\boxed{\binom{n}{2}=\frac{n(n-1)}{2}}$$

Indeed,

$$\binom{n}{2}=\frac{n!}{2!(n-2)!}=\frac{1\cdot2\cdot3\cdots n}{(1\cdot2)\cdot(1\cdot2\cdot3\cdots(n-2))}=\frac{(n-1)n}{2}$$

For example,

$$\binom{5}{2}=\frac{5\cdot4}{2}=10 \qquad \binom{10}{2}=\frac{10\cdot9}{2}=45$$

EXAMPLE 7

There are 10 people in a room. We choose 3 people out of them.

a) If we consider the 3 people as a group (there is no order)

there are 10C3 = 120 possible ways

b) If we arrange the 3 people in order

there are 10P3 = 720 possible ways

Particularly, for permutations the BOX-technique helps

$$\boxed{10}\ \boxed{9}\ \boxed{8} = 720 \text{ possible ways}$$

EXAMPLE 8 (LOTTO)

We choose 6 numbers out of 49. We clearly have combinations. The number of possible ways is

$$49C6 \quad \text{or} \quad \binom{49}{6}$$

That is

$$\frac{49!}{6!43!} = \frac{44\cdot45\cdot46\cdot47\cdot48\cdot49}{1\cdot2\cdot3\cdot4\cdot5\cdot6} = 13983816$$

This implies that if we choose 6 numbers the probability to win is 1 out of 13983816, that is almost 1 out of 14million!

EXAMPLE 9

There are 26 Latin letters (A, B, C, … , Z). When we construct words of 3 letters, the order of letters plays a role, thus we have permutations. However, it will be better to use the Box-technique instead of permutation formula nPr. The total number of these words is

$$\boxed{26}\ \boxed{26}\ \boxed{26}$$

that is 26^3 = **17576** words

Find

a) how many of them consist of different letters.
b) how many of them begin with A.
c) how many of them do not begin with A.
d) how many of them do not contain the letter F.
e) how many of them contain the letter F (at least once).

Solution

a)

$$\boxed{26}\ \boxed{25}\ \boxed{24}$$

That is (26)(25)(24)= **15600** words

b)

$$\underset{\textbf{reserved}}{\boxed{1}}\ \boxed{26}\ \boxed{26}$$

That is 26^2= **676** words

c)

$$\boxed{25}\ \boxed{26}\ \boxed{26}$$

That is $(25)(26)^2$= **16900** words

d) For each box we have only 25 choices (we exclude F)

$$\boxed{25}\ \boxed{25}\ \boxed{25}$$

That is 25^3 = **15625** words

e) We present two methods

Method A (analytical)

The word has one of the following forms

(x can be any letter except F)

Fxx	$\boxed{1}$	$\boxed{25}$	$\boxed{25}$	25^2=625
xFx	$\boxed{25}$	$\boxed{1}$	$\boxed{25}$	25^2=625
xxF	$\boxed{25}$	$\boxed{25}$	$\boxed{1}$	25^2=625
FFx	$\boxed{1}$	$\boxed{1}$	$\boxed{25}$	25
FxF	$\boxed{1}$	$\boxed{25}$	$\boxed{1}$	25
xFF	$\boxed{25}$	$\boxed{1}$	$\boxed{1}$	25
FFF	$\boxed{1}$	$\boxed{1}$	$\boxed{1}$	1

Totally: 625+625+625+25+25+25+1=**1951** words

Method B (smart way!)

It is sometimes more convenient to count exactly the opposite cases and exclude them from the total!

Here we exclude cases in d) from the total number of cases

$$17576-15625 = \mathbf{1951} \text{ words}$$

EXAMPLE 10

A school class consists of 30 students, 10 boys and 20 girls. We select a committee of 5 students. Clearly, the total number of all possible committees is

$$\binom{30}{5}$$

Find the number of ways to select a committee of 5 students if

a) the committee consists of 2 boys and 3 girls
b) the committee consists of boys only
c) the committee consists of students of the same gender
d) there are at most two boys in the committee

Solution

a) $\binom{10}{2}\binom{20}{3}$ ways

b) $\binom{10}{5}\binom{20}{0}=\binom{10}{5}$ ways

c) There are two cases. The committee contains only boys or only girls. Hence, the number of possible ways is $\binom{10}{5}+\binom{20}{5}$

d) There are three cases. The committee contains 0 or 1 or 2 boys (and 5,4,3 girls respectively). Hence the number of ways is

$$\binom{10}{0}\binom{20}{5}+\binom{10}{1}\binom{20}{4}+\binom{10}{2}\binom{20}{3}$$

- PROBABILITY AND COUNTING

The probability formula

$$P(A)=\frac{n(A)}{TOTAL}$$

seems to be very naïve. However, the estimation of both n(A) and TOTAL is not always trivial. It very often requires the counting techniques above. Roughly speaking, the "**number of ways**" for some particular case in counting becomes "**probability**", if we divide by the TOTAL number of all possible ways:

$$P(A)=\frac{n(A)}{TOTAL}\quad\begin{array}{l}\leftarrow \text{number of ways in question}\\ \leftarrow \text{number of all possible ways}\end{array}$$

We revisit two characteristic examples above.

EXAMPLE 11 (compare with EXAMPLE 9 above)

Find the probability that a word of 3 letters

a) consists of different letters.
b) begins with A.
c) does not begin with A.
d) does not contain the letter F.
e) contains the letter F (at least once).

Solution

We have seen that the total number of 3-letter words is 26^3

Hence,

a) P(different letters) = $\frac{(26)(25)(24)}{26^3} = \frac{150}{169}$

b) P(begins with A) = $\frac{(26)^2}{26^3} = \frac{1}{26}$

c) P(does not begin with A) = $\frac{(25)(26)^2}{26^3} = \frac{25}{26}$

d) P(does not contain the letter F) = $\frac{25^3}{26^3}$

e) P(contains the letter F) = $1-\frac{25^3}{26^3}$

EXAMPLE 12 (compare with EXAMPLE 10 above)

A school class consists of 30 students, 10 boys and 20 girls. Find the probability that a committee of 5 students

a) consists of 2 boys and 3 girls
b) consists of boys only
c) consists of students of the same gender
d) contains at most two boys

Solution

We have seen that the TOTAL number of possible 5-member committees is $\binom{30}{5}$

Hence, the answers are

a) $$\frac{\binom{10}{2}\binom{20}{3}}{\binom{30}{5}}$$

b) $$\frac{\binom{10}{5}\binom{20}{0}}{\binom{30}{5}}=\frac{\binom{10}{5}}{\binom{30}{5}}$$

c) $$\frac{\binom{10}{5}+\binom{20}{5}}{\binom{30}{5}}$$

d) $$\frac{\binom{10}{0}\binom{20}{5}+\binom{10}{1}\binom{20}{4}+\binom{10}{2}\binom{20}{3}}{\binom{30}{5}}$$